

THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Exploring epigenetic gene regulation using applied machine learning

Kashyap Chhatbar



THE UNIVERSITY of EDINBURGH
informatics

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computational Biology and Artificial Intelligence
2025

School of Informatics,
University of Edinburgh,
10 Crichton Street,
Edinburgh,
EH8 9AB
Kashyap Chhatbar ©2025

Declaration

I confirm that this thesis is my original work, composed solely by me. All contributions from other sources have been explicitly acknowledged within the text. This thesis has not been submitted, either in whole or in part, for any other degree or professional qualification except as stated.

Kashyap Chhatbar
March 2025

Acknowledgements

I am grateful to my line manager and supervisor, Adrian Bird, for his support, which enabled me to undertake this part-time PhD alongside my professional commitments. His guidance and perspective have been greatly appreciated.

I am profoundly grateful to my supervisor, Guido Sanguinetti, for his unparalleled patience and steadfast support throughout this journey. His willingness to accommodate my challenges, particularly in navigating statistical aspects of the research, has been instrumental in my growth as a scientist.

I would also like to extend my sincere thanks to the College of Science and Engineering for awarding me a scholarship to support my tuition fees, without which this endeavour would not have been possible. I would also like to thank my PhD committee, especially Duncan Sproul, for their feedback and constructive discussions that helped shape the direction of my research.

I am especially grateful to my friends and colleagues, Raphaël and Sara, for their insightful discussions, technical assistance, and camaraderie throughout this journey.

Finally, I can't thank my wife, Deepu, enough for her constant support, patience, and encouragement throughout this journey. From keeping me grounded during the most intense periods of writing to managing so many aspects of our lives while we balanced research and other professional commitments, she has been my constant source of strength. Her belief in me, even when I doubted myself, made all the difference, and I could not have done this without her.

Abstract

Computational biology plays a pivotal role in scientific discovery, particularly in molecular biology and biomedical research. Advances in high–throughput sequencing have enabled large–scale profiling of epigenomic marks across diverse cellular contexts, providing unprecedented opportunities for data–driven insights. Leveraging these vast datasets, machine learning and artificial intelligence (AI) approaches have become powerful tools for predicting gene expression and revealing molecular interactions. However, the correlative nature of these models often require rigorous validation and integration with experimental data. This thesis applies computational techniques, including linear and non–linear modelling, deep learning, and explainable AI (XAI), to investigate epigenetic mechanisms of gene regulation. By leveraging high–quality transcriptomic and epigenomic datasets, it reassesses longstanding hypotheses and quantifies the influence of chromatin–associated proteins on transcription.

One such well–characterized regulator, MeCP2, binds methylated CpG (mCG) di–nucleotides and mCAC tri–nucleotides, recruiting the NCoR complex to regulate gene expression in the brain. While this mechanism is well–established, alternative roles for MeCP2 have been proposed, including its involvement in RNA splicing and specific binding to tandemly repeated CA sequences. To explore MeCP2's role in RNA splicing, I analysed datasets spanning varying levels of MeCP2 and DNA methylation. Using bioinformatic approaches combined with linear and non–linear machine learning models, my findings revealed minimal splicing changes, with MeCP2 and DNA methylation accounting for only 5–10% of observed variation. Similarly, the proposed preferential targeting of CA repeats by MeCP2 was examined. I found that CA repeats undergo CAC methylation at levels typical of the genome, with no evidence supporting preferential binding. These results suggest that MeCP2–mediated regulation remains consistent regardless of CA repeat presence and is unlikely to play a broad role in splicing. This work reinforces MeCP2's primary function in methylation–dependent transcriptional regulation, while addressing critical gaps in the understanding of its proposed alternative functions.

While MeCP2 regulates transcription by invoking DNA methylation–dependent binding mechanisms, other chromatin–associated factors may employ different strategies. Mammalian genomes contain extensive regions of homogenous AT– or GC–rich DNA, conserved across species but with unclear functional significance. SALL4, previously shown to co-localise with AT–rich heterochromatin in cultured

cells, was found to target a wide range of AT–rich motifs based on analyses of *in vitro* and *in vivo* datasets from mouse embryonic stem cells (mESCs). Using data from SALL4 DNA–binding deficient mutants and SALL4 overexpression models in mESCs, I applied linear models to quantify the relationship between gene body AT content and the extent of gene expression changes. These results demonstrate that SALL4 represses genes in proportion to their AT content, underscoring the potential of DNA base composition as a biological signal for regulating gene expression.

My previous work focused on investigating the role of MeCP2 and SALL4 individually in transcriptional regulation. In the next phase, I aim to explore how chormatin–associated proteins function together to coordinate gene expression. Many modify histones post–translationally, generating epigenetic marks that downstream pathways interpret to coordinate transcriptional programs. Understanding this complex context–dependent interplay remains an open challenge. By generating deep learning models to predict RNA Pol–II occupancy from chromatin profiles in unperturbed mESC conditions, I applied Shapley Additive Explanations (SHAP), a widely used explainable AI (XAI) approach, to infer functional regulatory mechanisms. Genes ranked by SHAP importance accurately predicted direct perturbation targets, even from unperturbed data, reducing the reliance on costly experimental interventions. SHAP analysis further revealed cooperative roles of SET1A and ZC3H4 at promoters and distinct contributions of ZC3H4 at gene bodies in transcriptional regulation. Cross–dataset validation revealed a striking similarity in the direct targets of ZC3H4 and INTS11, suggesting potential regulatory convergence mediated through H3K4me3 and the SET1/COMPASS complex. This highlights SHAP's ability to uncover context–dependent transcriptional patterns and reveal previously unrecognized connections between regulatory pathways.

In summary, this thesis seeks to demonstrate the power of computational biology in dissecting transcriptional regulation by challenging assumptions and uncovering potential new regulatory paradigms. It refines our understanding of how chromatin–associated proteins interpret DNA sequence and epigenetic marks to control transcription. Finally, I propose that predictive modelling combined with XAI approaches like SHAP offers a scalable framework for predicting functional targets and infer regulatory mechanisms.

Lay Summary

Deoxyribonucleic acid (DNA) serves as the blueprint for life, containing the instructions for creating gene products like ribonucleic acid (RNA) and proteins. DNA is composed of four nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). The relative abundance of these bases within a gene, particularly the balance of A and T versus C and G, determines its base composition. In addition to base composition, linear DNA is highly organized and tightly wrapped around histone proteins, forming chromatin fibres. Cells don't turn all their genes on at once–they have special control systems that decide which genes should be active and when. One way this regulation happens is through chemical modifications to DNA, histone proteins, or both, which can influence how genes are switched on or off. Specialized proteins either bind directly to unmodified DNA or recognize and modify chemical marks on DNA and chromatin, ultimately controlling gene activity. This intricate system is known as the epigenetic regulation of gene expression. The overarching goal of this thesis is to explore how specialised proteins interact with unmodified DNA, chemically modified DNA, and chromatin to regulate gene expression. To achieve this, I employed data–driven research methods, leveraging machine learning and explainable artificial intelligence (XAI).

One important protein involved in this process is MeCP2, which recognizes chemically modified DNA, known as methylated DNA. MeCP2 ensures proper control of gene activity, especially in the brain. Non–functional versions of MeCP2 cause Rett Syndrome (RTT), a severe neurological disorder, making it a major focus of research. While MeCP2 is thought to have diverse roles, I explored two proposed functions: its influence on RNA splicing, a process that allows a single gene to produce different RNA products, and its binding to repetitive DNA sequences, regions of the genome where short DNA patterns appear many times in a row. Using machine learning, I found no evidence that MeCP2 or DNA methylation directly affects RNA splicing. Additionally, MeCP2's binding to repetitive DNA sequences is dictated by how closely they match its usual targets, with no unique regulatory role. By disproving these alternative functions, my findings refine the understanding of MeCP2's true role in gene regulation.

Another key protein I studied is SALL4, which helps maintain stem cells, i.e. cells that can develop into different types of tissues. By analysing smaller fragments of DNA sequences that are bound by SALL4, I discovered that SALL4 favours sequences rich in A and T nucleotides, rather than a specific sequence. Interestingly, the more A's and T's

a gene contains, the more likely SALL4 is to suppress its activity. SALL4's preference for A– and T–rich sequences highlights the importance of DNA base composition in shaping gene regulation. I propose that SALL4's ability to regulate genes across the genome is directly influenced by the fluctuating A and T content in DNA, as genes with higher A and T proportions are more likely to be suppressed. This suggests that variations in DNA base composition play a pivotal role in controlling gene expression and, ultimately, determining cell fate.

Each protein that regulates genes has a specific job, but figuring out how they work together as a team is challenging. One important process they control is transcription, where cells make RNA from DNA. This process relies on many proteins that attach to DNA and its packaging material, chromatin, to help turn genes on or off. To better understand these interactions, I trained advanced machine learning models to predict how RNA Polymerase II, the protein that makes RNA, behaves based on the presence of these other proteins near genes. Using explainable AI (XAI), I identified which proteins play the biggest role in controlling gene activity. To confirm my findings, I analysed data from experiments where these proteins had been removed, allowing me to see how gene activity changed in their absence. My AI–based predictions were highly accurate in forecasting how gene activity changes when key proteins are removed.

In conclusion, my research helps clarify how key proteins interact with chromatin to control gene activity. While I ruled out MeCP2's involvement in RNA splicing and its preference for repetitive DNA sequences, my findings revealed new ways in which proteins like SALL4 recognise DNA sequences and influence gene expression. Finally, by applying AI to biological data, I demonstrated that this approach can accurately predict how genes will behave, helping to deepen our understanding of gene regulation.

Contents

Abstract v

Lay Summary vii

Contents ix

List of Figures xi

Acronyms xiii

1 Introduction 1
- 1.1 Content of the thesis . . . 3
- 1.2 Contributions . . . 4

2 Background 5
- 2.1 A biologist's perspective . . . 5
  - 2.1.1 A brief history of DNA . . . 5
  - 2.1.2 From DNA to RNA . . . 6
  - 2.1.3 Splicing . . . 6
  - 2.1.4 Transcription factors . . . 8
  - 2.1.5 Chromatin . . . 9
  - 2.1.6 Histone modifications . . . 10
  - 2.1.7 Chromatin mediated transcriptional control . . . 12
  - 2.1.8 Co-regulator complexes . . . 13
  - 2.1.9 DNA methylation . . . 16
  - 2.1.10 Applications of genome engineering . . . 18
- 2.2 High-throughput sequencing and its applications . . . 20
  - 2.2.1 Quantifying the process of transcription . . . 21
  - 2.2.2 Detecting protein interactions with DNA . . . 22
  - 2.2.3 Probing chromatin accessibility . . . 23
  - 2.2.4 Finding the fifth base . . . 23
- 2.3 An informaticist's perspective . . . 24
  - 2.3.1 Foundations of *bio*–informatics . . . 24
  - 2.3.2 Quantification of biological signal . . . 24
  - 2.3.3 Applied machine learning techniques . . . 25
  - 2.3.4 Interpretable machine learning techniques . . . 31

3 Research questions 35
- 3.1 Molecular roles for MeCP2 . . . 35
- 3.2 Short DNA motifs as regulators of transcription . . . 37
- 3.3 Regulation of transcription by chromatin-associated proteins . . . 38
  - 3.3.1 The role of PRC1 and PRC2 in transcription . . . 39
  - 3.3.2 The role of INTS11 in transcription . . . 40
  - 3.3.3 The role of ZC3H4 in transcription . . . 41

4 Reassessing the molecular roles of MeCP2 **43**
4.1 MeCP2's role in influencing alternative splicing . . . 43
4.1.1 Introduction . . . 43
4.1.2 Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing . . . 43
4.1.3 Conclusion . . . 58
4.2 The Impact of CA Repeats on MeCP2 Function . . . 58
4.2.1 Introduction . . . 58
4.2.2 A critique of the hypothesis that CA repeats are primary targets of neuronal MeCP2 . . . 58
4.2.3 Conclusion . . . 68

5 Influence of DNA base composition and SALL4 on gene regulation **69**
5.1 Introduction . . . 69
5.2 SALL4 controls cell fate in response to base composition . . . 69
5.3 Conclusion . . . 93

6 Interpreting epigenetic regulation of gene expression **95**
6.1 Introduction . . . 95
6.2 Modeling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies . . . 95
6.3 Conclusion . . . 121

7 Discussion **123**
7.1 Summary of contributions . . . 123
7.1.1 Reevaluating MeCP2's Role in Splicing and Binding . . . 123
7.1.2 AT-rich DNA as a regulatory signal . . . 124
7.1.3 Inferring regulatory interactions in transcriptional regulation . . . 124
7.2 Outlook . . . 125
7.2.1 Is MeCP2 multi-functional? . . . 125
7.2.2 SALL4's role in transcription . . . 126
7.2.3 Enhancing SHAP analysis for transcriptional regulation . . . 126

A Additional methods and source code **129**
Supplementary material for Section 4.1.2 . . . 129
Supplementary material for Section 5.2 . . . 134
Supplementary material for Section 6.2 . . . 145
Source code for bioinformatics workflows and models . . . 152

References **153**

List of Figures

1.1 Data explosion 1
1.2 ENCODE assays by year 2

2.1 The mechanisms of gene expression 7
2.2 Transcription factor (TF) and motif status for each TF family 8
2.3 Mechanisms of TF effector domain action to affect gene expression . . . 9
2.4 Heirarchical chromatin-folding model 10
2.5 Schematic representation of histone modifications 11
2.6 Models for coordinated recruitment of co-repressors and co-activators . 15
2.7 NCoR and NuRD complex recruitment by MeCP2 and SALL4 16
2.8 Chemical-genetics enables the temporal analysis of transcription 20
2.9 Development of NGS platforms 21
2.10 Overview of Selected HTS Applications 21
2.11 Basic building block of an MLP 29
2.12 Representation of a feedforward neural network 29

Acronyms

5caC 5-carboxylcytosine 17
5fC 5-formylcytosine 17
5hmC 5-hydroxymethylcytosine 17
5mC 5-methylcytosine 16, 17
AI Artificial Intelligence 2, 123
AID Auxin-Inducible Degron 39
ALKBH2 alpha-ketoglutarate dependent dioxygenase homologu 2 16
ATP Adenosine Triphosphate 14
CFP1 CXXC Finger Protein 1 12
CGIs CpG Islands 12, 17, 18, 37, 40, 41, 69, 125, 126
CHD Chromodomain Helicase DNA-binding 15
CUT&Tag Cleavage Under Targets and Tagmentation 125
DNA Deoxyribonucleic acid 1, 5
DNMT DNA Methyltransferase 16–18
dTAG degradation tag 39
EMBL European Molecular Biology Laboratory 24
eRNA enhancer RNA 41
FFNN feedforward neural network 29
H2Aub Mono-uniquitylation of Histone 2A 12
H2Bub Mono-uniquitylation of Histone 2B 12
H3K27me3 Trimethylation of Lysine 27 in Histone 3 11
H3K36me2 Dimethylation of Lysine 36 in Histone 3 11
H3K36me3 Trimethylation of Lysine 36 in Histone 3 11
H3K4me3 Trimethylation of Lysine 4 in Histone 3 11, 12
H3K79me2 Dimethylation of Lysine 79 in Histone 3 12
H3K9me3 Trimethylation of Lysine 9 in Histone 3 11
HAT Histone Acetyltransferase 12
HDAC Histone Deacetylases 12
HMT Histone Methyltransferases 18
HP1 Heterochromatin Protein 1 11
INTS Integrator Complex Subunit 40
KDMs Lysine Demethylases 11
KMTs Lysine Methyltransferases 11
MBD Methyl-CpG-Binding Domain 17
MeCP2 Methyl CpG Binding Protein 2 15, 123
ML Machine Learning 2, 3
MLP Multilayer Perceptron 29
mRNA Messenger Ribonucleic Acid 1, 5, 6
MSE Mean Squared Error 25, 30

NCoR Nuclear Receptor Co-repressor 15
ncRNA non-coding RNA 41
NuRD Nucleosome Remodeling and Deacetylase 15, 38, 39, 126
OLS Ordinary Least Squares 30
Pol–II RNA Polymerase II 6–8, 11, 13, 40–42, 95, 125
PPI Protein-Protein Interactions 8
PRC1 Polycomb Repressive Complex-1 12, 39
PRC2 Polycomb Repressive Complex-2 11, 39
PROMPT promoter upstream transcripts 41
PTMs Post-Translational Modifications 9, 10
RNA RiboNucleic Acid 5, 6
RTT Rett Syndrome 18, 35, 125, 126
SALL Spalt-like 15, 126
SETD SET domain containing protein 11, 41
SGD Stochastic Gradient Descent 30
SHAP SHapley Additive exPlanations 25, 31, 95, 123, 126
snRNPs Small Nuclear Ribonucleoproteins 6
TET Ten-Eleven Translocation methylcytosine dioxygenase 17
TFs Transcription Factors 6, 8, 17, 126
UHRF1 Ubiquitin-like-containing PHD and RING finger domain 1 17
WDR WD repeat domain 41
XAI Explainable Artificial Intelligence 2, 3, 25, 95
ZC3H Zinc Finger CCCH-Type Containing 41

Chapter 1

Introduction

Notable advancements in biology are often a result of reductionist approaches. The strategy of breaking complex systems into their fundamental components, has worked so far in uncovering biological principles. Significant contributions like the discovery that Deoxyribonucleic acid (DNA) is the genetic material, the one gene–one enzyme hypothesis, the central dogma of molecular biology, the discovery of Messenger Ribonucleic Acid (mRNA) emerged from reductionist approaches. To drive them, there is a great demand for technologies which deliver precise quantification of experimental data. As a welcome consequence, modern biology has embraced big data (Figure 1.1) and emerged as a highly quantitative science. Moving forward, a key challenge to advance quantitative bioscience is to integrate approaches from several disciplines. Computational biology is an evolving field that applies mathematical modelling, statistical analysis and machine learning. Therefore, it has become a cornerstone of this interdisciplinary effort (Markowetz 2017).

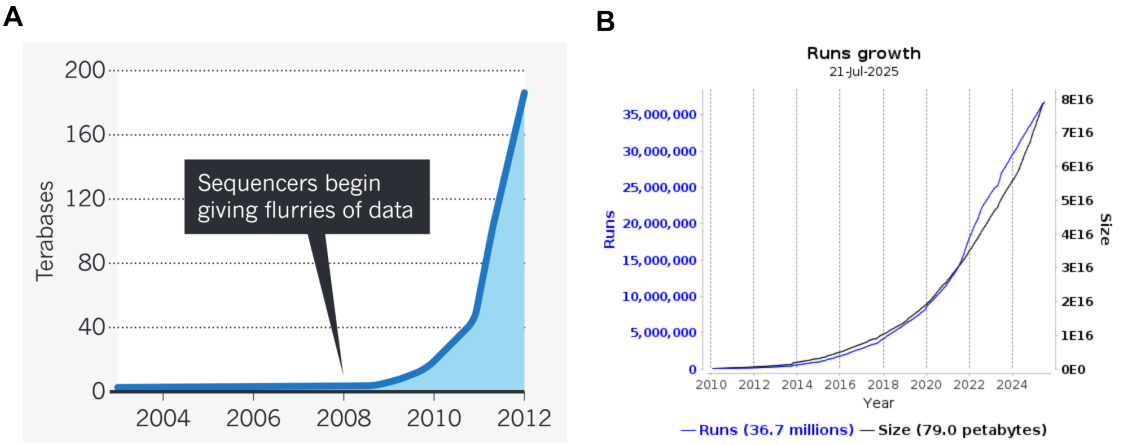


Figure 1.1. Data explosion. A) Figure from Marx 2013. The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size. B) A more updated version of the graph downloaded from https://www.ebi.ac.uk/ena/browser/about/statistics

Biological data are inherently heterogeneous, reflecting the diversity of experimental techniques used to explore gene regulation (Figure 1.2). This regulation involves multiple layers of control driven by interactions among DNA-binding proteins, chromatin modifications, and DNA methylation. Large scale epigenomic assays provide genome-wide profiles (The ENCODE Project Consortium 2012), facilitating large-scale analysis and modelling. These datasets not only reveal

co-occurrence patterns but also generate testable hypotheses, which can be experimentally validated through genetic perturbations. Advances such as degron technology (Jaeger and Winter 2021), which enables rapid degradation of specific proteins, present exciting opportunities to dissect causal regulatory relationships. As vast datasets accumulate from different developmental stages and experimental conditions, standardised analytical approaches are critical for extracting meaningful insights.

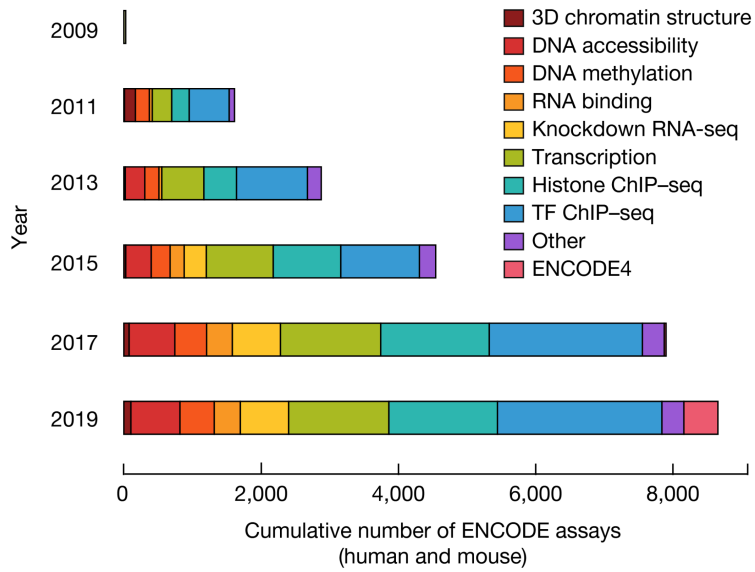


Figure 1.2. ENCODE assays by year. Figure from The ENCODE Project Consortium et al. 2020. Accumulations of assays over the three phases of ENCODE.

The sequencing of the human genome marked the dawn of the genomics era, where the sheer scale of data necessitated advanced computational solutions (Pabinger et al. 2014). Modern workflows allow data to be processed and standardized, opening new frontiers in Artificial Intelligence (AI) and Machine Learning (ML) applications (Angermueller et al. 2016). However, the same models that provide unparalleled predictive power often act as black boxes, offering little insight into their decision-making processes. Yet, in biological sciences, accuracy alone is not enough–interpretability is equally essential. The ability to explain and justify a model's predictions fosters trust, ensures reproducibility, and enables researchers to uncover causal relationships, transforming machine learning from a mere predictive tool into a powerful instrument for scientific discovery (Molnar 2022).

Through the application of ML and Explainable Artificial Intelligence (XAI), this thesis seeks to refine our understanding of epigenetic gene regulation. By ensuring model transparency and interpretability, this work showcases the use of data–driven

research to enable mechanistic exploration.

1.1 Content of the thesis

In Chapter 2, I provide the foundational concepts of DNA methylation, chromatin, chromatin–associated proteins, gene regulation, applied ML and XAI. Chapter 3 builds upon this by reviewing the literature and defining the key research questions addressed in this thesis. Appendix A points to supplementary methods detailing experimental procedures and an online resource containing the source code for the computational methodologies used to investigate these questions, providing a framework for the analyses conducted in subsequent chapters. The core chapters of this thesis are based on the following publications:

- Chapter 4 comprises the following peer-reviewed journal articles:

  > Kashyap Chhatbar et al. (Oct. 13, 2020). "Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing". In: *PLOS Genetics* 16.10. Ed. by Dirk Schübeler, e1009087. DOI: 10.1371_journal.pgen.1009087

  > Kashyap Chhatbar et al. (Dec. 2022). "A critique of the hypothesis that CA repeats are primary targets of neuronal MeCP2". In: *Life Science Alliance* 5.12, e202201522. DOI: 10.26508_lsa.202201522

- Chapter 5 comprises the following peer-reviewed journal articles:

  > Raphaël Pantier, Kashyap Chhatbar, et al. (Feb. 2021). "SALL4 controls cell fate in response to DNA base composition". In: *Molecular Cell* 81.4, 845–858.e8. DOI: 10.1016_j.molcel.2020.11.046

- Chapter 6 comprises a preprint undergoing revision following first round of peer-review:

  > Kashyap Chhatbar et al. (July 15, 2025). *Modeling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies.* DOI: 10.1101_2025.01.30.635704

Finally, I conclude this thesis by summarising the key findings from every chapter and discuss future outlook in Chapter 7.

1.2 Contributions

As the primary author of three papers and co-first author of one paper included in this thesis, I played a central role in designing the research, analysing the results, and writing the manuscripts. Throughout this process, my advisors, Guido Sanguinetti and Adrian Bird, provided invaluable guidance–helping refine experimental designs, offering methodological insights, and shaping the manuscripts with their detailed feedback. Raphaël Pantier and Timo Quante, co-authors of the paper presented in Chapter 5, led the research by designing the experiments, conducting laboratory work, analysing the results, and drafting the manuscript. John Connelly, co-author of the paper presented in Section 4.2.2, co-led the research by designing the *in-vitro* experiments, conducting laboratory work, analysing the results, and drafting the manuscript.

Chapter 2

Background

This chapter focuses on introducing the background, from the biologist and the informaticist perspective, highlighting key concepts relevant to this thesis. Selective omissions are intentional for clarity and conciseness, without prejudice.

2.1 A biologist's perspective

2.1.1 A brief history of DNA

The polymer DNA is the carrier of instructions for the development and reproduction of all living organisms and many viruses. The discovery of "nuclein" (Miescher 1871) and prediction about nucleic acids by Friedrich Miescher in 1869 were not widely appreciated by the scientific community for nearly half a century. Although the groundwork had been laid, significant contributions from Phoebus Levene, mainly, discovery of the order of three major components phosphate-sugar-nucleobase forming a single nucleotide (Levene 1919) was the catalyst for further research. It was not until Oswald Avery's experiments, demonstrating that DNA could transfer genetic information (McCarty and Avery 1946), that its significance as the molecule of heredity was recognised. Erwin Chargaff expanded on Levene's work and reached a major conclusion, now known as Chargaff's rule, that the total amount of purines (adenine and guanine) and the total amount of pyrimidines (cytosine and thymine) are usually nearly equal (Chargaff 1950). All these discoveries culminated in the famous double-helical model of DNA made possible by contributions from Francis Crick, James Watson, Rosalind Franklin and Maurice Wilkins (Franklin and Gosling 1953; Watson and Crick 1953; Wilkins et al. 1953). After the acceptance of the genetic role of DNA as a "working hypothesis", a quest began to explain how the transfer of information occurred inside the cell. During the seminal lecture by Francis Crick in 1957, published later (Crick 1958), he addressed the central problems of molecular biology – gene action and nucleic acid synthesis. This work, along with the discovery of mRNA (Brenner et al. 1961; Gros et al. 1961; Jacob and Monod 1961), built upon research by many groups helped cement a new fundamental view of how life inside a cell works, "DNA makes RNA makes protein".

2.1.2 From DNA to RNA

The mechanisms of gene expression

Gene expression starts with the conversion of DNA into RNA (Kornberg 2007). In eukaryotes, Transcription Factors (TFs) play a key role by guiding RNA polymerases to promoters–the start sites of genes. In eukaryotes, RNA Polymerase II (Pol–II) functions to transcribe DNA into precursors of mRNA. Once this process starts, especially transcription of protein-coding genes, Pol–II moves along the DNA and synthesises a nascent RNA. Soon after transcriptional initiation, the nascent RNA undergoes 5' capping, which protects it from degradation and facilitates downstream processes. The transition from initiation to elongation is tightly regulated by elongation factors (Uptain et al. 1997), ensuring efficient transcript extension. Within the gene, coding sequences (exons) are interrupted by non-coding sequences (introns), which must be removed through pre-mRNA splicing to generate a functional transcript (Figure 2.1A). Upon reaching the gene's end, transcription termination is coupled with RNA cleavage and polyadenylation, stabilising the transcript and influencing its nuclear export and translation (Orphanides and Reinberg 2002; Proudfoot et al. 2002; Reed and Hurt 2002). While classical biochemical approaches have studied these steps in isolation, recent advances demonstrate that transcription and splicing are tightly interconnected, often occurring co-transcriptionally (Figure 2.1B). This shift in perspective highlights the dynamic coordination of gene expression, where regulatory mechanisms act across multiple steps rather than as discrete, independent events.

2.1.3 Splicing

Constitutive splicing (Figure 2.1C) is the fundamental process by which introns are removed and exons are joined together to form a mature mRNA transcript in eukaryotic cells. This process is highly efficient and relies on the recognition of strong, conserved splice sites by the spliceosome, a complex of Small Nuclear Ribonucleoproteins (snRNPs) and associated proteins (Marasco and Kornblihtt 2023). Alternative splicing (Figure 2.1C), in contrast, allows a single gene to produce multiple mRNA isoforms by selectively including or excluding specific exons, utilizing alternative splice sites, or retaining introns. This process significantly expands proteomic diversity, as per the prediction from Walter Gilbert (Gilbert 1978).

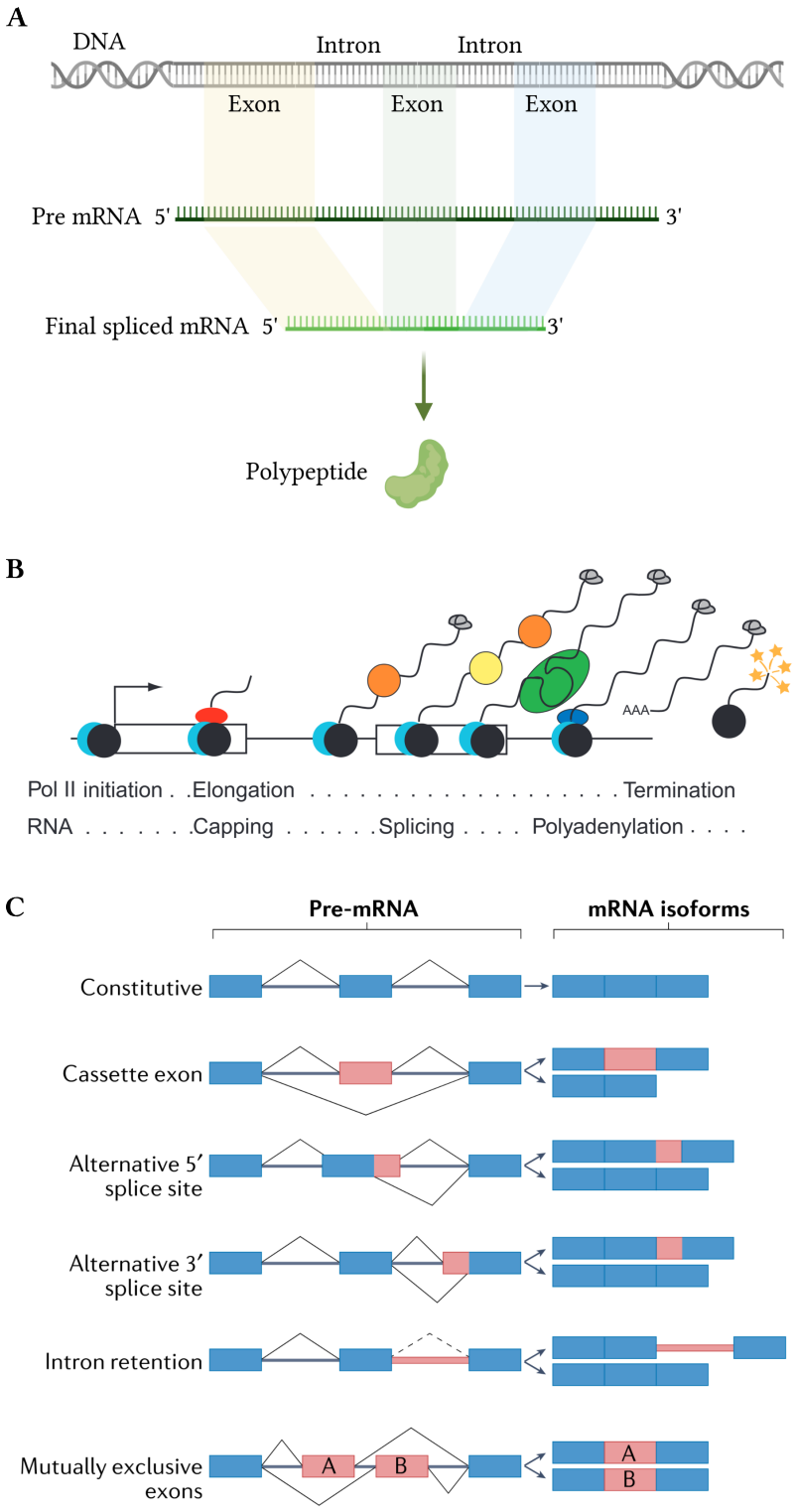


Figure 2.1. The mechanisms of gene expression A) Structure of a eukaryotic gene. Adapted from genome.gov/genetics-glossary/Gene B) Co-transcriptional pre-mRNA processing indicating Pol–II (black ball), polyadenylation factors (blue and dark blue balls), capping enzyme (red oval), splicing factors (orange and yellow balls), spliceosome (green oval). Taken from Neugebauer 2002. C) Alternative splicing. Figure taken from Marasco and Kornblihtt 2023.

2.1.4 Transcription factors

The term TFs, introduced in Section 2.1.2, broadly refers to proteins involved in transcription and gene regulation. However, its modern definition is more precise, encompassing only proteins that (1) bind DNA and (2) regulate transcription (Lambert et al. 2018). While sequence-specific DNA binding is often assumed to be a defining characteristic, studies have shown that hundreds of TFs lack strict sequence specificity (Jolma et al. 2013) (Figure 2.2). TFs often function cooperatively, either by assisting one another in DNA binding, modifying chromatin accessibility, or mediating Protein-Protein Interactions (PPI) to influence transcription. TFs exert their effects through diverse mechanisms–some directly recruit Pol–II to promoters, while others operate by engaging cofactors that modulate transcriptional activity. This is possible because of dedicated effector domains within TFs (Figure 2.3). However, a considerable number of TFs lack well-defined effector domains, instead relying on low-complexity regions and short functional motifs for cofactor interactions (Erijman et al. 2020). Traditionally, TFs have been classified as either activators or repressors, but this binary view is increasingly challenged, likely because of the ability of certain TFs to recruit multiple cofactors (Soto et al. 2022). Many TFs exhibit context-dependent functions, where their regulatory effects vary depending on the local chromatin landscape, DNA sequence, and cofactor availability (Frietze and Farnham 2011; Reiter et al. 2017).

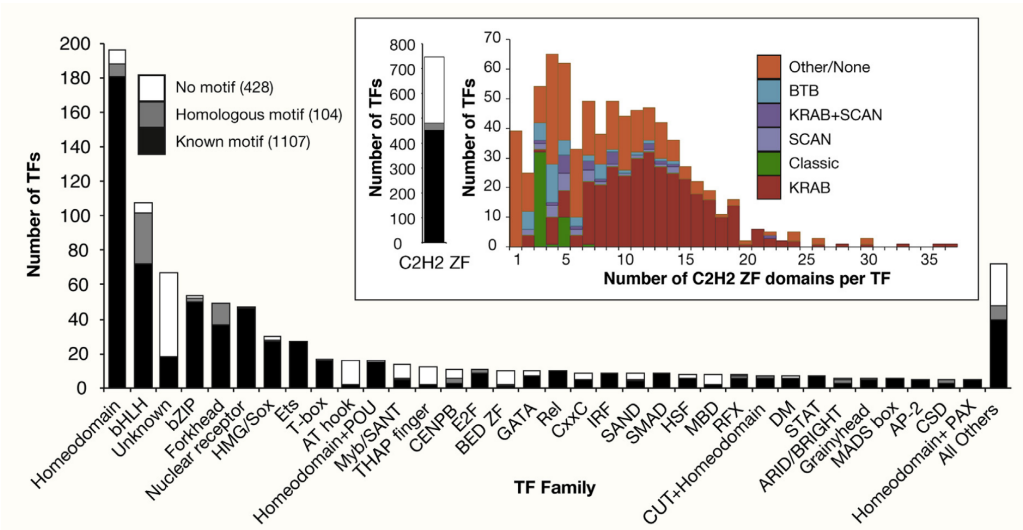


Figure 2.2. Transcription factor (TF) and motif status for each TF family highlighting a large number of TFs showing no known motif (white colour in stacked bars). Figure taken from Lambert et al. 2018.

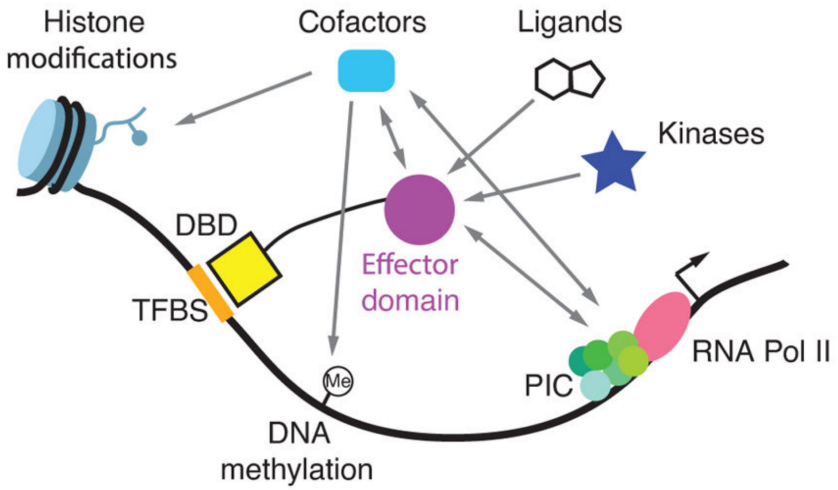


Figure 2.3. Mechanisms of TF effector domain action to affect gene expression. Effector domains regulate gene expression by interacting with cofactors and the preinitiation complex, modifying histones, altering DNA methylation, and responding to ligands or post-translational modifications. Taken from Soto et al. 2022

2.1.5 Chromatin

The DNA in eukaryotic cells is not freely floating but is packaged into a highly organized nucleoprotein complex called chromatin. The fundamental unit of chromatin is the nucleosome, consisting of 146 base pairs (bp) of DNA wrapped around a core of histone proteins (H2A, H2B, H3, and H4) (Luger et al. 1997). This packaging serves a dual function–compacting the genome within the nucleus (Figure 2.4) while simultaneously regulating gene expression. Chromatin exists in two primary states: heterochromatin, which is highly compact and transcriptionally repressed (Allshire and Madhani 2018), and euchromatin, which is more accessible and associated with active transcription (Meneghini et al. 2003). Beyond its structural role, chromatin is dynamically regulated through histone modifications, which occur on the flexible N-terminal tails of histone proteins that extend outward from the nucleosome. These tails contain lysine residues, which serve as sites for various Post-Translational Modifications (PTMs) such as acetylation, methylation, phosphorylation, and ubiquitination (Millán-Zambrano et al. 2022).

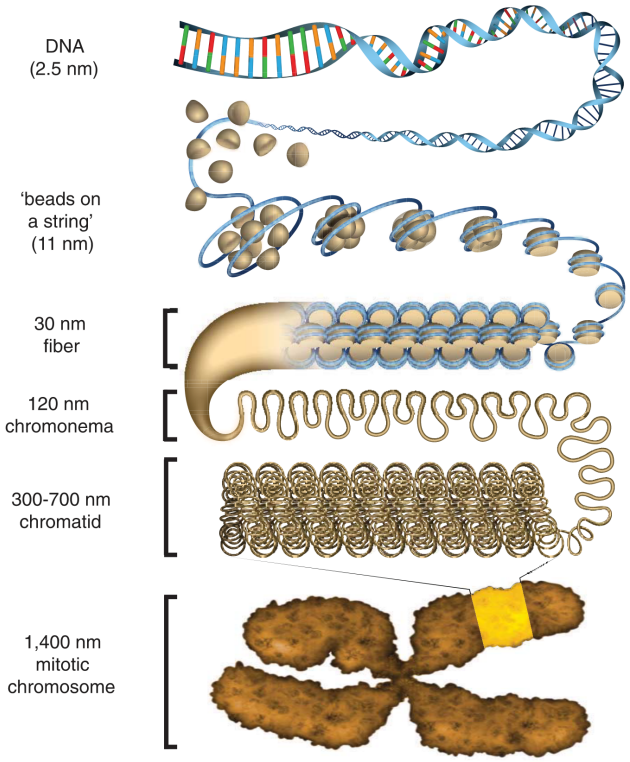


Figure 2.4. Hierarchical chromatin-folding model. Figure taken from Ou et al. 2017

2.1.6 Histone modifications

Histone PTMs (Figure 2.5) influence gene regulation through diverse mechanisms, acting both directly and indirectly. In many cases, histone modifications appear to be a consequence of upstream biological processes rather than the cause of transcriptional changes (Millán-Zambrano et al. 2022). Whether histone modifications serves as a primary driver of gene expression or merely functions as an intermediary within broader regulatory networks remains an ongoing debate (Henikoff and Shilatifard 2011). However, recent evidence suggests that certain histone marks actively drive or repress gene expression, reinforcing the idea that histone modifications can play instructive roles in transcriptional regulation (Policarpi et al. 2024).

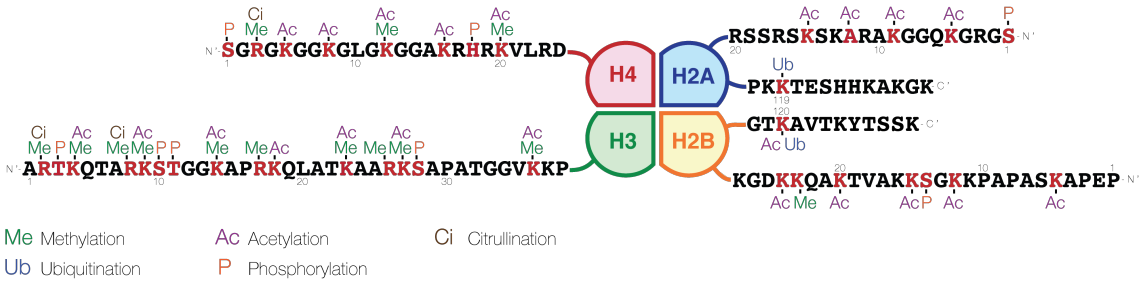


Figure 2.5. Schematic representation of histone modifications. Figure taken from Mariuswalter 2015 which was adapted from Rodríguez-Paredes and Esteller 2011.

Methylation

One of the most well-characterized histone modifications is Trimethylation of Lysine 4 in Histone 3 (H3K4me3), a mark of active transcription that is highly enriched at the promoters of active genes in eukaryotes (Vermeulen et al. 2007). It typically occurs mutually exclusive to DNA methylation (Section 2.1.9), reinforcing distinct transcriptional states (Hughes et al. 2020). Another key methylation mark, H3K36me3, is strongly associated with transcriptional elongation due to the recruitment of the SET domain containing protein 2 (SETD2) by elongating Pol–II (Bannister et al. 2005; Kizer et al. 2005; Vakoc et al. 2006). Beyond activation-associated marks, H3K27me3 and H3K9me3 are key repressive histone modifications linked to gene silencing. H3K27me3, catalyzed by Polycomb Repressive Complex-2 (PRC2), is essential for silencing developmental genes and is inherited across cell divisions (Schuettengruber and Cavalli 2009). It can also accumulate following transcriptional repression, highlighting its role in maintaining gene silencing (Simon and Kingston 2009). H3K9me3, a marker of constitutive heterochromatin, is enriched in transcriptionally silent regions and bound by Heterochromatin Protein 1 (HP1), which facilitates chromatin compaction and further repression (Allshire and Madhani 2018). Loss of H3K9me3 has been linked to the derepression of repetitive elements and some genes, though its effects are cell-type dependent (Nicetto et al. 2019). In humans, H3K36me2 is enriched across megabase-scale genomic domains, where it interacts with DNA methylation to shape the intergenic methylation landscape (Weinberg et al. 2019). The deposition and removal of lysine methylation on both histone and non-histone proteins is mediated by "writers" Lysine Methyltransferases (KMTs) and "erasers" Lysine Demethylases (KDMs). The human genome encodes over 100 KMTs, and studies suggest that more than 1,000 proteins in the human proteome are methylated, highlighting the extensive regulatory potential of lysine methylation (Husmann and Gozani 2019).

Acetylation

Histone acetylation is widely associated with active transcription, yet its role is more nuanced than a simple activation switch. While budding yeast lacking H3 or H4 N-terminal tails remain viable, suggesting that histone modifications are not strictly essential for transcription (Ling et al. 1996), global acetylation levels influence gene expression in a dose-dependent manner (Dion et al. 2005; Kurdistani et al. 2004). The relationship between histone acetylation and gene expression is further complicated by findings in human T cells, where both Histone Acetyltransferase (HAT) and Histone Deacetylases (HDAC) are enriched at actively transcribed genes, and HDACs are not preferentially found at silent genes (Wang et al. 2009). This suggests that acetylation functions dynamically, facilitating nucleosome turnover and chromatin accessibility rather than directly activating or repressing transcription. In Drosophila, acetylated nucleosomes are rapidly replaced, enriched in ATP-dependent chromatin remodelers, and associated with DNAseI hypersensitivity, supporting the idea that acetylation–deacetylation cycles maintain chromatin accessibility (Kharchenko et al. 2011). Further evidence from mutational studies of H4 acetylation sites indicates that gene expression levels correlate with the total number of acetylated lysines rather than its presence at specific residues, reinforcing the concept that acetylation modulates nucleosome interactions rather than serving as a binary regulatory mark (Dion et al. 2005; Megee et al. 1995; Shogren-Knaak et al. 2006).

Ubiquitylation

Histone ubiquitylation is a post-translational modification that regulates both transcriptional activation and repression, as well as DNA damage repair. Unlike acetylation and methylation, which involve small chemical groups, ubiquitylation entails the covalent attachment of ubiquitin, a 76-amino acid polypeptide, to lysine residues. Mono-uniquitylation of Histone 2A (H2Aub), catalyzed by Polycomb Repressive Complex-1 (PRC1), is linked to gene silencing (Schuettengruber and Cavalli 2009), Mono-uniquitylation of Histone 2B (H2Bub) is associated with transcription elongation and recruitment of histone methyltransferases involved in H3K4me3 and H3K79me2 deposition (Vitaliano-Prunier et al. 2008).

2.1.7 Chromatin mediated transcriptional control

SET1 and H3K4me3 regulation of transcription

The SET1/COMPASS complex catalyses the trimethylation of histone H3 at lysine 4 (H3K4me3), a modification traditionally associated with transcriptionally active CpG Islands (CGIs) promoters (Hughes et al. 2020) via the conserved subunit CXXC Finger Protein 1 (CFP1) (Clouaire et al. 2012). It is proposed that H3K4me3 deposition further

reinforces SET1 occupancy and enables recruitment of additional histone reader proteins to modify chromatin and influence polymerase activity (Klose et al. 2013). SET1 complexes contain a protein called WDR82 that directly interacts Pol–II and could provide alternative mechanisms to influence transcription (Lee et al. 2010). The SET1/COMPASS complex exhibits evolutionary conservation from yeast to humans, witih different family members (SET1A/B, MLL1-4) displaying specialised functions (Shilatifard 2012).

INTS11 and the Integrator complex

The Integrator complex, which includes the subunit INTS11, plays a crucial role in regulating transcription by interacting with RNA polymerase II (Pol II) and facilitating the processing of non-coding RNAs. At protein-coding genes, INTS11 mediates Pol–II pause-release through endonucleolytic cleavage of nascent transcripts, allowing for the recruitment of elongation-competent Pol–II complexes (Baillat et al. 2005; Wagner et al. 2023). The Integrator complex functions as part of a larger regulatory network that includes the Integrator-PP2A complex (INTAC), which combines RNA endonuclease and protein phosphatase activities to control Pol–II dynamics. INTAC dephosphorylates the Pol–II C-terminal domain at Ser2, Ser5, and Ser7, providing coordinated regulation of transcriptional elongation. The architectural separation of the endonuclease (INTS11) and phosphatase modules within INTAC allows for distinct regulatory functions while maintaining integrated control over transcription (Zheng et al. 2020).

ZC3H4 and the Restrictor complex

The Restrictor complex represents a critical quality control mechanism that distinguishes productive from spurious transcription across the genome ZC3H4, an RNA-binding protein, partners with WDR82 and ARS2 to form the Restrictor complex, which suppresses transcription of non-coding RNAs including promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs), and antisense transcripts. The complex preferentially localizes at transcription start sites of active genes through direct interactions between ZC3H4 and WDR82 subunits with the Ser5-phosphorylated Pol–II C-terminal domain (Estell et al. 2023, 2021; Estell and West 2025).

2.1.8 Co-regulator complexes

Gene regulation is orchestrated by the coordinated recruitment of co-activators and co-repressors, which dynamically modify chromatin to regulate transcription. The classical model suggests that TFs recruit these complexes in a stable manner, with co-activators such as HATs promoting transcription by modifying chromatin structure, while co-repressors such as HDAC-containing complexes impose transcriptional

silencing (Figure 2.6A). However, this model does not fully account for the dynamic nature of transcriptional regulation, as growing evidence shows that gene activation and repression are not static states but are instead highly responsive to environmental and developmental signals (Perissi et al. 2010).

To address this limitation, the cyclical model proposes that transcription factors, co-activators, and co-repressors continuously cycle on and off chromatin, leading to a competitive exchange between activation and repression (Figure 2.6B). This turnover allows genes to remain poised for rapid transcriptional changes, enabling fine-tuned responses to cellular cues (Bracken et al. 2006). More recent combinatorial models attempt to unify these mechanisms, proposing that co-activators and co-repressors do not function in isolation but instead operate in interdependent regulatory networks (Figure 2.6C). Rather than a binary switch, these models suggest that gene expression is dictated by the dynamic balance of co-regulator interactions (Wang et al. 2009). Advances in high-throughput sequencing and genome-wide profiling have begun to test these models, revealing gene- and context-specific regulatory patterns that further challenge the traditional view of transcriptional regulation (Barski et al. 2007; Mikkelsen et al. 2007; The ENCODE Project Consortium 2012).

Chromatin remodellers

Chromatin remodelers are ATP-dependent protein complexes that regulate nucleosome positioning, composition, and density, thereby influencing transcription, DNA replication, and repair. By using Adenosine Triphosphate (ATP) hydrolysis, these remodelers alter histone-DNA interactions, facilitating nucleosome sliding, eviction, or histone exchange (Clapier and Cairns 2009). Remodelers are classified into distinct families, each with specialized roles in chromatin dynamics. These complexes recognize both modified and unmodified nucleosomes, allowing them to target specific chromatin regions based on histone modifications or DNA accessibility. Many remodelers also interact with co-regulators, linking chromatin remodeling to broader regulatory networks. They frequently recruit HATs or HDACs, integrating chromatin remodeling with histone acetylation or deacetylation to promote either an open or closed chromatin state, respectively (Legube and Trouche 2003; Perissi et al. 2010; Roth et al. 2001; Xu et al. 1999). Through this interplay, co-regulators influence gene expression not only through histone modifications but also by modulating chromatin accessibility via remodeler activity.

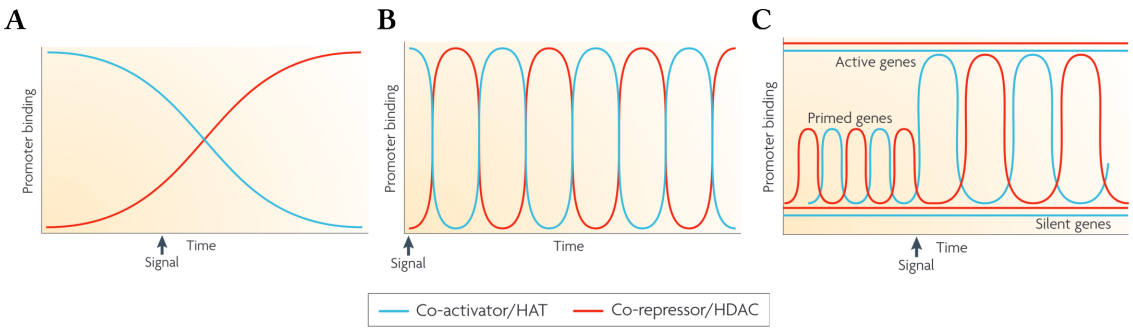


Figure 2.6. Models for coordinated recruitment of co-repressors and co-activators. Taken from Perissi et al. 2010

NuRD complex

The Nucleosome Remodeling and Deacetylase (NuRD) complex, which integrates chromatin remodelling and histone deacetylation, was first identified through its HDAC activity and association with transcriptional repression (Tong et al. 1998; Wade et al. 1998; Xue et al. 1998; Zhang et al. 1998). Central to this complex are the Chromodomain Helicase DNA-binding (CHD) proteins 3 and 4 (CHD3/4), which belong to the CHD family of ATP-dependent chromatin remodellers. These proteins contribute to nucleosome repositioning and are crucial for gene silencing, differentiation, and DNA damage response (Clapier and Cairns 2009; Denslow and Wade 2007; Lai and Wade 2011). Beyond its role in gene silencing, NuRD is involved in lineage commitment, ensuring that developmental and tissue-specific gene expression programs are tightly regulated (Bowen et al. 2004). Furthermore, NuRD dynamically interacts with DNA methylation machinery (Wade et al. 1999; Zhang et al. 1999), linking chromatin remodelling, histone deacetylation, and epigenetic regulation to maintain genomic stability. The conserved N-terminus of Spalt-like (SALL) proteins is known to recruit NuRD and repress transcription (Kiefer et al. 2002; Lauberth and Rauchman 2006; Lu et al. 2009) (Figure 2.7).

NCoR complex

The Nuclear Receptor Co-repressor (NCoR) complexes are key regulators of transcriptional repression, primarily functioning through their association with HDACs. Initially identified as co-repressors for nuclear hormone receptors (Guenther et al. 2000; Li 2000), these complexes have since been recognized as broader regulators of chromatin remodelling and gene silencing. Both NCOR1 and NCOR2 serve as scaffolds for the recruitment of HDAC3, facilitating histone deacetylation and chromatin compaction to repress transcription (Perissi et al. 2010). The NCoR complex is recruited by Methyl CpG Binding Protein 2 (MeCP2) that mediates transcriptional repression by linking DNA methylation with histone deacetylation (Nan et al. 1998)

(Figure 2.7). This interaction was found to be disrupted by Rett syndrome-associated mutations, which abolish MeCP2's ability to recruit NCoR (Lyst et al. 2013).

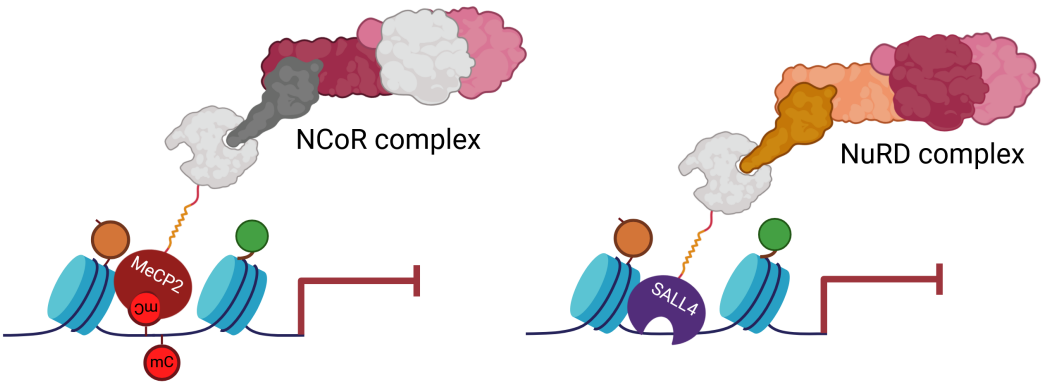


Figure 2.7. NCoR and NuRD complex recruitment by MeCP2 and SALL4

2.1.9 DNA methylation

This is a concise overview of DNA methylation relevant to this thesis. For a deeper understanding, readers are encouraged to explore these excellent reviews (Greenberg and Bourc'his 2019; Jaenisch and Bird 2003; Jones 2012; Klose and Bird 2006).

Tracing the origins

5-methylcytosine (5mC) was first synthesised *in vitro* (Wheeler 1904) and later identified in living organisms, with its presence confirmed in *Mycobacterium tuberculosis* (Johnson and Coghill 1925). Subsequent studies using paper chromatography led to its detection as "epi-cytosine" (Hotchkiss 1948), a finding later validated as 5mC, solidifying its role as a natural DNA modification (Wyatt 1950, 1951). Despite its ancient evolutionary roots in bacteria, 5mC has been independently lost across multiple eukaryotic lineages (Zemach and Zilberman 2010). Notably, model organisms such as *Drosophila melanogaster, Caenorhabditis elegans* and certain yeasts exhibit little to no 5mC. The inherent mutagenicity of 5mC stems from its tendency to undergo spontaneous deamination, leading to C → T transitions (Holliday and Grigg 1993; Rideout et al. 1990), a process that has significantly shaped genome composition (Bird and Taggart 1980). As a consequence, mammalian genomes contain approximately five-fold fewer CpG dinucleotides than expected based on their nucleotide composition. Despite its mutagenic nature, the persistence of DNA methylation in many eukaryotes suggests an evolutionary balance. The co-evolution of DNA methylation writers, DNA Methyltransferase (DNMT) enzymes, alongside alkylation repair enzymes like alpha-ketoglutarate dependent dioxygenase homologu 2 (ALKBH2), may have provided a compensatory mechanism, allowing DNA

methylation to be maintained without excessive mutational burden (Rošić et al. 2018).

Writers and erasers

The establishment and regulation of DNA methylation encompass distinct phases: de novo DNA methylation, maintenance and demethylation. In mammals, the enzymes DNMT3A and DNMT3B catalyse de novo methylation by transferring a methyl group to cytosine residues using their conserved DNMT domain (Okano et al. 1999, 1998). The maintenance of CpG methylation during DNA replication is primarily carried out by DNMT1, which interacts with a protein called Ubiquitin-like-containing PHD and RING finger domain 1 (UHRF1) (Nishiyama et al. 2013; Qin et al. 2015). UHRF1 detects hemimethylated CpG dinucleotides at replication forks and relieves the auto-inhibitory conformation of DNMT1, thereby ensuring methylation on the daughter DNA strand (Ishiyama et al. 2017). Active removal of methylation marks is mediated by Ten-Eleven Translocation methylcytosine dioxygenase (TET) dioxygenases, which oxidise 5mC to 5-hydroxymethylcytosine (5hmC) (Kriaucionis and Heintz 2009), 5-formylcytosine (5fC) (Ito et al. 2010) and 5-carboxylcytosine (5caC) (Ito et al. 2011). These oxidised intermediates facilitates DNA demethylation through replication-dependent dilution (Hashimoto et al. 2012; Otani et al. 2013).

Reading, recruiting and repressing

Early *in vitro* and *in vivo* studies established a fundamental role of 5-methylcytosine (5mC) in gene repression (Holliday and Pugh 1975; Riggs 1975), a conclusion further supported by the strong anti-correlation between promoter 5mC levels and transcriptional activity (Weber et al. 2007). While CpG methylation is a ubiquitous feature across all somatic cells, non-CpG methylation (mCH, where H = A, T, or C) is largely restricted to pluripotent stem cells and neurons, hinting at cell-type-specific regulatory roles (He and Ecker 2015). A prevailing model for 5mC-mediated repression Figure 2.7 suggests that methylation physically hinders TFs binding (Stadler et al. 2011), thereby impeding transcriptional activation. Additionally, 5mC acts as a docking site for Methyl-CpG-Binding Domain (MBD) proteins (Hendrich and Bird 1998; Meehan et al. 1989), which recruit chromatin-modifying complexes that reinforce a repressive chromatin state (Nan et al. 1998; Ng et al. 1999).

Hotspots across genome

Perhaps the most recognized term associated with DNA methylation is the CGIs, regions dense in CpG nucleotides yet usually devoid of 5mC (Bird et al. 1985). These relatively small (1̃kb on average) genomic elements are predominantly found at promoters, and being unmethylated, explains their evolutionary conservation despite their CpG-rich composition. In contrast, gene bodies are CpG-poor and extensively methylated, making

it a major source of germline mutations and a key contributor to oncogenic mutations in somatic cells (Holliday and Grigg 1993; Rideout et al. 1990). However, an intriguing subset of orphan CGIs resides within gene bodies, functioning as alternative promoters with potential regulatory roles in development (Illingworth et al. 2010). Beyond CpG methylation, accumulating evidence highlights the presence of non-CpG methylation, particularly in mammalian brain, suggesting a specialised role in neuronal function (De Mendoza et al. 2021).

Neurological disorders

Epigenetic regulation, particularly through DNA methylation and histone modifications, plays a crucial role in neurodevelopment. These mechanisms influence neuronal differentiation and cognitive function, and their disruption is increasingly linked to neurological disorders (Reichard and Zimmer-Bensch 2021). A well-studied neurological disorder linked to DNA methylation is Rett Syndrome (RTT), caused by mutations in MeCP2 (Amir et al. 1999). Extensive evidence suggests that its inability to bind methylated DNA (mCG and mCAC) or recruit the NCoR co-repressor complex disrupts gene regulation, contributing to this severe disorder (Lyst et al. 2013). Conversely, excess MeCP2 results in MeCP2 duplication syndrome, underscoring the importance of balanced methylation-mediated repression (Del Gaudio et al. 2006; Van Esch et al. 2005). Beyond this, DNA methylation machinery is also implicated through targeted recruitment of DNMT enzymes via chromatin modifications. The HMT, NSD1 directly recruits DNMT3A to chromatin, shaping the methylome (Hamagami et al. 2023; Weinberg et al. 2019). Mutations in DNMT3A cause Tatton-Brown-Rahman syndrome (TBRS) (Childhood Overgrowth Consortium et al. 2014), an intellectual disability disorder, while NSD1 mutations lead to Sotos syndrome (Kurotaki et al. 2002), another neurological condition. Given their shared molecular mechanisms, these disorders exhibit significant phenotypic overlap, further highlighting the interplay between histone modification and DNA methylation in neurodevelopment (Hamagami et al. 2023; Weinberg et al. 2019).

2.1.10 Applications of genome engineering

Precise and targeted genome modification is essential for determining the functions of specific genes. Researchers can selectively inactivate a gene of interest or introduce mutations that mimic disease-causing variants to evaluate their phenotypic consequences in cells or whole organisms. Historically, gene targeting relied on homologous recombination, in which two similar DNA strands exchange nucleotide sequences, but its efficiency in higher eukaryotic cells is extremely low (1 in 10^6 to 1 in 10^7) (Kim and Kim 2014). More recently, programmable nucleases have revolutionized

genome editing by generating site-specific DNA double-strand breaks (DSBs), which significantly enhance the efficiency of homologous recombination (Bibikova et al. 2002; Rouet et al. 1994). This rapidly evolving technology reached its peak with the discovery that the clustered regularly interspaced short palindromic repeat (CRISPR)-Cas adaptive immunity system in bacteria could be repurposed for precise genome editing, offering a simpler and more efficient alternative to previous methods (Jinek et al. 2012; Sternberg et al. 2014; Wiedenheft et al. 2012). Mechanistically, the CRISPR system consists of two key components: a single-stranded guide RNA (sgRNA) and an endonuclease. The sgRNA is designed to complement the target DNA sequence with high specificity, guiding the endonuclease to the correct locus, where it induces a DSB. The resulting break activates endogenous DNA repair pathways, enabling precise genome modifications through the introduction of small insertions or deletions.

However, genetic manipulation typically occurs at the DNA or RNA level, while cellular phenotypes are primarily driven by changes in protein abundance, activity, and interactions. As a result, the effects of genetic modifications often manifest slowly and are largely limited to equilibrium measurements, making it difficult to distinguish direct consequences from secondary or tertiary effects (Figure 2.8). To overcome these limitations, an increasing number of studies leverage pharmacological protein-level perturbations, enabling a more immediate and mechanistic understanding of cellular function (Banaszynski et al. 2006; Holland et al. 2012). This approach utilizes the same genome engineering tools, but instead of modifying the gene itself, the goal is to alter the resulting protein. Specifically, the protein of interest (POI) is fused to a degron tag. Upon exposure to an inducing ligand, the degron-fused protein is targeted for destruction via the ubiquitin-proteasome pathway, allowing for acute depletion with precise temporal control. Rapid target protein depletion is particularly valuable for studying highly dynamic nuclear processes such as transcriptional regulation (Jaeger and Winter 2021). Ligand-induced degron systems have emerged as powerful tools in this context, allowing researchers to achieve acute protein depletion with precise temporal control (Dobrinić et al. 2021; Hughes et al. 2023; Wang et al. 2023).

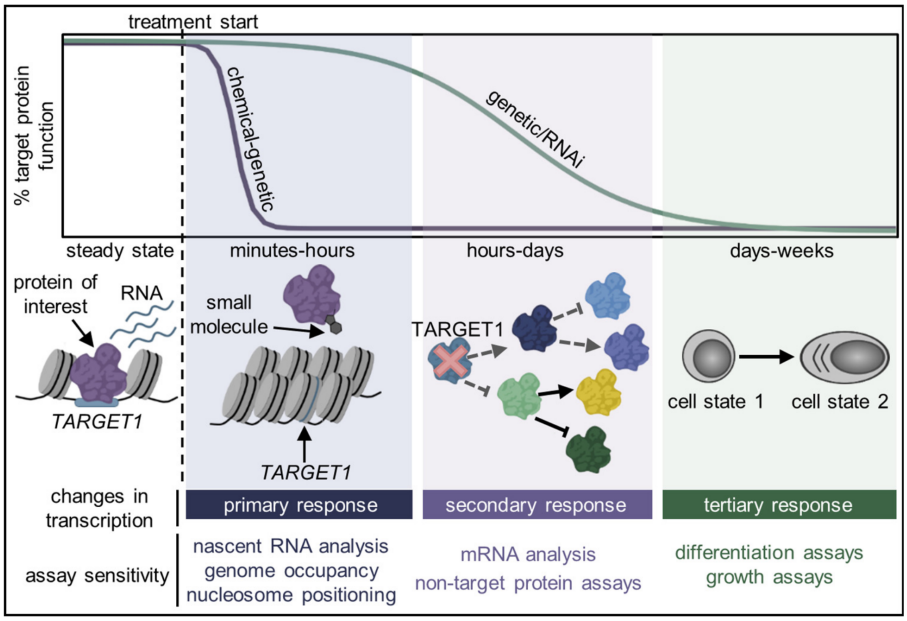


Figure 2.8. Chemical-genetics enable the temporal analysis of transcription. Traditional genetic perturbation methods, such as knockouts or RNAi knockdowns, take days to weeks to establish transcriptional dysregulation, making it difficult to separate primary regulatory effects from secondary responses. In contrast, chemical-genetic approaches can rapidly disrupt protein function within minutes, allowing for precise temporal control. When combined with nascent RNA analysis, these methods enable a clearer distinction between immediate transcriptional responses and downstream effects. Figure taken from Layden et al. 2024.

2.2 High-throughput sequencing and its applications

DNA sequencing has been instrumental in advancing our understanding of genetics, enabling researchers to decode entire genomes with increasing speed and accuracy. The first widely adopted sequencing method, Sanger sequencing (Sanger and Coulson 1975) relied on selective chain termination using dideoxynucleotides, followed by electrophoretic separation and fluorescence-based detection. While Sanger sequencing became the gold standard for sequencing smaller DNA fragments, its scalability was a major limitation for large-scale projects. Despite these challenges, it was the primary method used to generate the first draft of the human genome, requiring over a decade of effort and billions of dollars in funding. After the completion of the Human Genome Project in 2001, there was a need for more efficient sequencing technologies (Schloss 2008). The advent of high-throughput sequencing (HTS) using updated technologies, dubbed the Next-Generation Sequencing (NGS) showed a flurry of development from many companies Figure 2.9. The applications of HTS extend far beyond genomics and whole-genome sequencing. It has become a cornerstone of various "omics" studies,

including transcriptomics and epigenomics Figure 2.10.

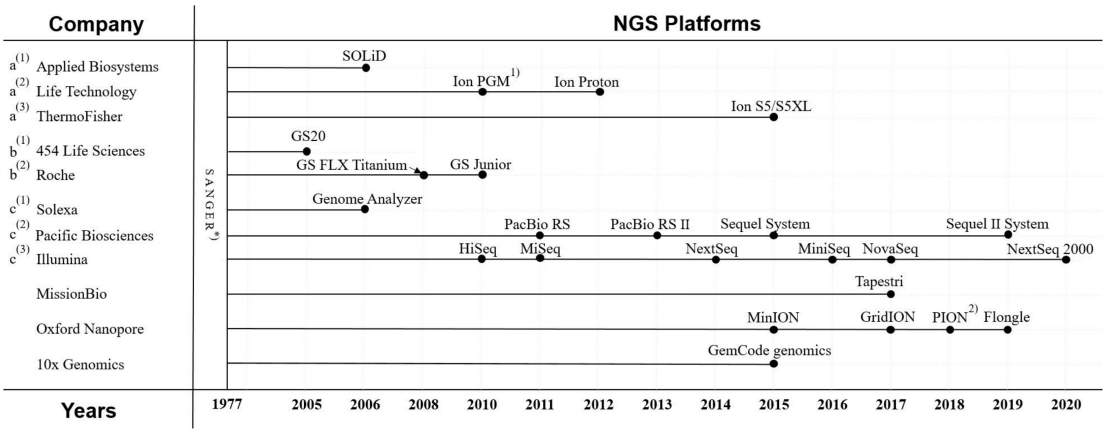


Figure 2.9. Development of NGS platforms. Taken from Kim et al. 2020

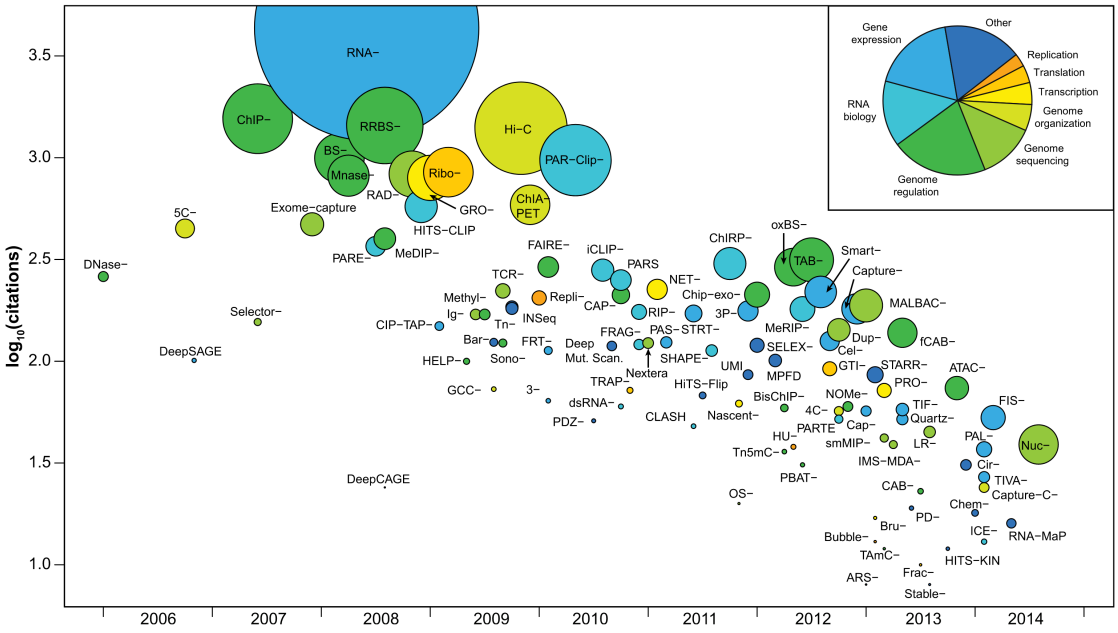


Figure 2.10. Overview of Selected HTS Applications. Publication date of a representative article describing a method versus the number of citations that the article received. Taken from Reuter et al. 2015

.

2.2.1 Quantifying the process of transcription

Gene expression is measured by quantifying transcriptional activity across all genes, with RNA sequencing (RNA-seq) being the most widely used method. RNA-seq involves RNA extraction, enrichment or depletion of specific unwanted RNA types, cDNA library preparation, and high-throughput sequencing (Stark et al. 2019). Initially

developed to measure steady-state mRNA levels, it has since been adapted to capture newly synthesized or nascent RNA. Nascent RNA-seq methods can be broadly categorized into three groups: (1) run-on methods, which rely on the incorporation of nucleotide analogues, (2) Pol–II immunoprecipitation-based methods, and (3) metabolic pulse-labelling using the nucleotide analogue 4-thiouridine (4sU). A limitation of pulse-labelling methods requiring long labelling times is that most of the transcript becomes labelled, reducing sensitivity. By selectively capturing only the newly transcribed RNA at the 3' end of transcripts near RNA polymerase II, transient transcriptome sequencing (TT-seq) enhances sensitivity (Schwalb et al. 2016). Similarly, methods such as thiol(SH)-linked alkylation for metabolic sequencing of RNA (SLAM-seq) allow quantification by introducing an additional alkylation step to modify 4sU, enabling direct detection through mutational analysis, thereby allowing simultaneous measurement of both nascent and mature RNA (Herzog et al. 2017).

2.2.2 Detecting protein interactions with DNA

Mapping protein–DNA interactions and epigenetic marks is essential for understanding gene regulation. The primary method for studying protein–DNA binding *in vivo* is chromatin immunoprecipitation (ChIP), in which antibodies selectively enrich DNA fragments bound to specific proteins or post-translationally modified histone tails. When combined with high-throughput sequencing (ChIP-seq), this technique enables genome-wide mapping of these interactions (Furey 2012). However, ChIP-seq has notable limitations, particularly those arising from cross-linking artifacts and background noise (Meyer and Liu 2014; Teytelman et al. 2013). To overcome these challenges, Cleavage Under Targets & Release Using Nuclease (CUT&RUN) was developed, utilizing a fusion of protein A/G with micrococcal nuclease (pAG-MNase) to selectively cleave antibody-bound chromatin without the need for crosslinking (Skene and Henikoff 2017). Because this process occurs within intact cells or nuclei, clipped chromatin fragments freely diffuse into solution, allowing for easy removal of unbound chromatin and significantly reducing background signal. An advancement over CUT&RUN, Cleavage Under Targets & Tagmentation (CUT&Tag) incorporates a fusion of protein A/G with Tn5 transposase, which simultaneously fragments DNA and inserts sequencing adapters at antibody-bound sites (Kaya-Okur et al. 2019). Compared to CUT&RUN and ChIP-seq, CUT&Tag offers even lower background and higher sensitivity, enabling efficient profiling from low cell numbers or even single cells, with streamlined library preparation and enhanced signal-to-noise ratios. In comparative assessments, CUT&Tag captured about half the peaks detected by ENCODE ChIP-seq for certain

histone acetylation marks such as H3K27ac, indicating that this method may miss regions identified by established techniques (Abbasova et al. 2025).

2.2.3 Probing chromatin accessibility

Mammalian genomes contain a diverse array of functional elements beyond coding regions. Many of these elements are more accessible to DNase I digestion, and the resulting DNase I hypersensitive sites have been identified as promoters, enhancers, and other key regulatory regions (Boyle et al. 2008). As discussed in Section 2.1.5, most genomic DNA is wrapped around nucleosomes with varying degrees of accessibility. The developers of Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) hypothesized that Tn5 transposase, a prokaryotic enzyme, could efficiently target and insert sequencing adaptors into open chromatin regions. Compared to DNase-seq, ATAC-seq provides a simpler and more efficient workflow, simultaneously fragmenting and tagging accessible chromatin in a single step, eliminating the need for complex multistep procedures (Buenrostro et al. 2013). Additionally, ATAC-seq requires significantly fewer cells (as few as 500 nuclei) while maintaining a high signal-to-noise ratio and strong correlation with DNase-seq, making it a powerful tool for chromatin accessibility profiling.

2.2.4 Finding the fifth base

As introduced in Section 2.1.9, and often described as the elusive and dynamic "fifth letter" of the "book of life", 5-methylcytosine (5mC) is commonly referred to as DNA methylation (Lister and Ecker 2009). The gold-standard method for high-resolution detection of DNA methylation is bisulfite (BS) conversion. In this process, genomic DNA is treated with sodium bisulfite, which selectively converts unmethylated cytosines into uracil, while methylated cytosines (5mC) remain unchanged. Following bisulfite treatment, synthesis of the complementary strand and high-throughput sequencing (HTS) enables determination of the methylation status at each cytosine by identifying whether the sequenced base appears as thymine (unmethylated) or cytosine (methylated) (Frommer et al. 1992). Additionally, sequencing independent template molecules from the same locus provides a digital measurement of cytosine methylation frequency, offering precise quantification of this epigenetic modification.

2.3 An informaticist's perspective

2.3.1 Foundations of *bio*–informatics

The origins of bioinformatics trace back to the 1960s and 1970s, coinciding with the collection of genetic sequence information. After the collection of genetic sequences, development of an efficient method to arrange these sequences and identify regions of similarity was required. The Needleman-Wunsch algorithm (Needleman and Wunsch 1970) was one of the first to enable this comparison using the principles of dynamic programming (Bellman 2003). By the 1980s, GenBank and European Molecular Biology Laboratory (EMBL) DNA database were established to store nucleotide sequences, marking a turning point in biological data organization (Smith 1990). These repositories grew rapidly with the advent of the first-generation DNA sequencing technologies, setting the stage for large-scale genomic projects especially by automation. This explosion of data led to the development of heuristic search methods, most notably the BLAST algorithm (Altschul et al. 1994), which revolutionized sequence alignment by significantly reducing computational time while maintaining sensitivity. This advance made it feasible to scan entire genomes considerably faster. Over time, bioinformatics expanded beyond sequence alignment to include gene prediction, structural biology, functional annotation, and systems biology, establishing itself as an indispensable discipline at the intersection of biology, mathematics, and computer science.

2.3.2 Quantification of biological signal

As outlined in Section 2.2, HTS technologies enable genome-wide quantification of transcription, chromatin interactions, and epigenetic modifications. Extracting meaningful biological insights from HTS data requires robust computational processing, quality control, and standardised analytical pipelines. The sequencing-based assays relevant for this thesis –RNA-seq, ChIP-seq, ATAC-seq, and whole-genome bisulfite sequencing (WGBS-seq)– each provide distinct yet complementary information about gene regulation.

RNA-seq quantifies gene expression by measuring transcript abundance, while ChIP-seq maps protein-DNA interactions, such as transcription factor binding and histone modifications. ATAC-seq identifies accessible chromatin regions by leveraging Tn5 transposase activity, and WGBS-seq profiles DNA methylation at single-base resolution using bisulfite conversion. These sequencing modalities rely on standardised computational workflows to ensure consistency and accuracy. Influenced by the ENCODE data processing pipelines, I have developed standardised bioinformatics

pipelines in Snakemake (Köster and Rahmann 2012) to ensure reproducibility and accuracy in data processing. Computational workflows are uploaded to the online resource highlighted in Chapter A. Briefly, data processing steps start with raw sequencing reads in FASTQ format which undergo quality control, alignment to a reference genome, and subsequent quantification using tools specific to each data type. Each modality captures distinct but interconnected biological signals, allowing for the systematic interrogation of gene regulatory mechanisms using computational modelling.

2.3.3 Applied machine learning techniques

Machine learning has become an essential tool in computational biology, enabling researchers to model complex biological systems without the need for explicit mechanistic assumptions. At its core, modelling in machine learning follows a standard workflow that includes data pre-processing, feature extraction, model training, and evaluation. Data pre-processing involves cleaning and normalising raw biological data, handling missing values, and ensuring consistency across datasets. Feature extraction, *e.g.* selects biological signal value at different functional elements of the genome, which serve as inputs to the model. Once features are defined, the model is trained using supervised or unsupervised learning approaches—supervised learning utilises labelled datasets, whereas unsupervised learning identifies patterns in unlabelled data. Model evaluation then assesses performance using metrics like Mean Squared Error (MSE) or coefficient of determination (R^2) for regression tasks or classification accuracy for categorical predictions. Traditional machine learning approaches include linear regression and random forests. A major advancement in recent years has been the adoption of deep learning models to extract meaningful representations from biological data (Angermueller et al. 2016). Unlike traditional models, deep learning frameworks can learn hierarchical feature representations, reducing the need for manual feature engineering. The growing availability of large-scale biological datasets, such as those generated by HTS, has further driven the adoption of machine learning techniques.

While deep learning models achieve high predictive accuracy, they are often criticized as black-box models due to their complex and multi-layered architectures. This makes it difficult to understand how individual features contribute to predictions. Recent advances in XAI methods, such as SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) and attention-based mechanisms, have made it possible to interpret model outputs (Molnar 2022). SHAP assigns importance scores to each feature, enabling researchers to quantify the contribution of specific genomic or epigenetic features to a given prediction. These methods help bridge the gap between

high predictive power and biological interpretability, allowing machine learning models to generate not only accurate but also biologically meaningful insights.

Linear Regression

The goal of regression is to estimate the value of one or more continuous *target* variables y based on a given d-dimensional vector x of *input* variables. This relationship is not perfectly deterministic but is instead influenced by an additional component, known as the *disturbance term* or *error variable* ε. ε represents an unobserved random factor that introduces variability or "noise" into the association between the *target* variable and the *input* variables, which are also referred to as *regressors*. This noise accounts for discrepancies between the predicted and actual values, capturing the effects of unknown or unmeasured influences on the target variable.

Linear regression model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{id} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

where $i = 1, \ldots, n$ and $\top$ is the transpose such that $\mathbf{x}_i^\top \boldsymbol{\beta}$ is the inner product between vectors $\mathbf{x}_i$ and $\boldsymbol{\beta}$. In simple linear regression, $d = 1$ and $\boldsymbol{\beta}$ is either a $d + 1$-dimensional *parameter vector* if an intercept β_0 is included in the model, otherwise $\boldsymbol{\beta}$ is d-dimensional. This can be stacked together and also written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

For the detailed notation, please refer below

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

This is the classical linear model for regression and the method to solve it using least squares was individually developed by Adrien-Marie Legendre and Carl Friedrich Gauss (Plackett 1972; Stigler 1981). Generalised linear models (GLMs), proposed by (Nelder and Wedderburn 1972) extend the classical linear model for regression and has found great application in computational biology (Nakagawa and Schielzeth 2010).

Linear regression is one of the most widely used models for both prediction and interpretation due to its straightforward mathematical structure. The model expresses the target variable as a weighted sum of input features, where each feature's contribution is represented by a coefficient. These coefficients quantify the relationship between each predictor and the outcome, allowing for direct interpretation of how changes in one variable influence the prediction while holding others constant. Importantly, in a linear regression model, the effect of a predictor is fixed across all data points, meaning that the influence of a change in a predictor is the same regardless of the values of other variables. This absence of context dependence makes linear regression particularly interpretable, but also limits its ability to capture interactions or variable-dependent effects. The coefficients $\boldsymbol{\beta}$ indicate the effect size of all features. For numerical features, β_i represents the expected change in the target variable for a one-unit increase in x_i, assuming all other features remain constant. For categorical features, the coefficient describes the difference in the target value relative to a reference category. The magnitude of a coefficient reflects the strength of its influence, while the sign, + or −, determines whether the relationship is positive or negative, respectively. The effect size can be further contextualized using standardized coefficients, which scale all predictors to have unit variance, allowing for comparison across different variables. While linear regression provides an interpretable framework, careful attention must be given to model assumptions and potential confounding relationships between features.

Linear regression models, while interpretable, have several limitations that can affect their reliability. They assume linearity between features and the target, struggle with multicollinearity, and require homoscedasticity (constant variance of errors). Violations of independence in observations or the presence of outliers can further distort estimates. Additionally, linear models treat feature effects as fixed, making them less effective for capturing interactions or non-linear relationships. These weaknesses highlight the need for alternative approaches when dealing with complex data.

Random forest regression

Predictive models like linear regression are global models, meaning they assume a single predictive relationship holds across the entire data space. However, when dealing with high-dimensional data where features interact in complex, non-linear ways, fitting a single global model can be both difficult and, when successful, often unintuitive.

Prediction trees offer a more flexible alternative by partitioning the data space into regions where simpler relationships can be learned. The core idea is straightforward: predictions are made by growing a binary decision tree. At each internal node, a

splitting rule is applied based on an input feature, directing the observation to either the left or right sub-branch. This process continues until a leaf node is reached, at which point the final prediction is made. Instead of enforcing a single global equation, prediction trees adapt to the data by recursively dividing it into smaller, more homogeneous subsets. The prediction at each leaf is then obtained by averaging (in regression) or taking the majority class (in classification) of the training observations that reach that node. This approach allows the model to capture complex interactions between features while remaining intuitive and interpretable, as the decision rules can be directly examined. However, individual trees can be prone to over-fitting, which motivates the use of ensemble methods like random forests to improve stability and predictive accuracy. Unlike linear regression, which assumes a linear relationship between the *target* variables y and the *input* variables x, random forests can capture complex, non-linear dependencies without requiring explicit feature transformations.

A random forest consists of an ensemble of B independent regression trees, where each tree f_b ($b = 1, \dots, B$) is trained on a bootstrapped subset of the training data. The prediction for a new input $\mathbf{x}'$ is obtained by averaging the outputs of all individual trees:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x}')$$

Each regression tree f_b is constructed by recursively partitioning the input space into disjoint regions, minimizing the variance within each partition.

The random forest algorithm was first introduced by Breiman (Breiman 2001) and has since become a widely used technique in machine learning, with extensive applications in computational biology (Chen and Ishwaran 2012). Notably, it has been employed for gene expression analysis, regulatory motif prediction, and protein function classification, demonstrating its versatility in handling complex biological data (Libbrecht and Noble 2015).

Multilayer Perceptron

The term *deep neural network* originates from efforts to develop mathematical models of information processing in biological systems. Over time, it has been broadly applied to a variety of models with different architectures and learning mechanisms. In essence, a neural network refers to any differentiable function that can be expressed as a computational graph, where the "nodes" represent primitive operations and the "edges" correspond to numerical data in the form of vectors, matrices, or tensors. Conceptually, this graph can be structured as a sequence of layers, with the term *deep*

referring to networks with multiple such layers (Rosenblatt 1962; Rumelhart et al. 1986).

At the core of neural networks is the artificial "neuron" (Figure 2.11), a computational unit that transforms its input through a weighted sum followed by a nonlinear activation function. These neurons form the fundamental building blocks of larger architectures, including the Multilayer Perceptron (MLP). MLP, also known as a feedforward neural network (FFNN), is one of the simplest forms of deep learning models. It consists of a series of L linear transformations interleaved with element wise non-linearities, allowing it to learn complex mappings between inputs and outputs.



Figure 2.11. Basic building block of an MLP. a) The output Y is a weighted combination of the inputs X, where the weights vector is denoted by W. b) Alternative depiction of the neuron's behaviour where the bias term can be emulated by defining $w_N = b$ and $X_N = 1$. Taken from Murphy 2023.

Mathematically, an MLP with L layers is expressed as:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \mathbf{W}_L\varphi_L\left(\mathbf{W}_{L-1}\varphi_{L-1}\left(\cdots\varphi_1\left(\mathbf{W}_1\boldsymbol{x}\right)\cdots\right)\right)$$

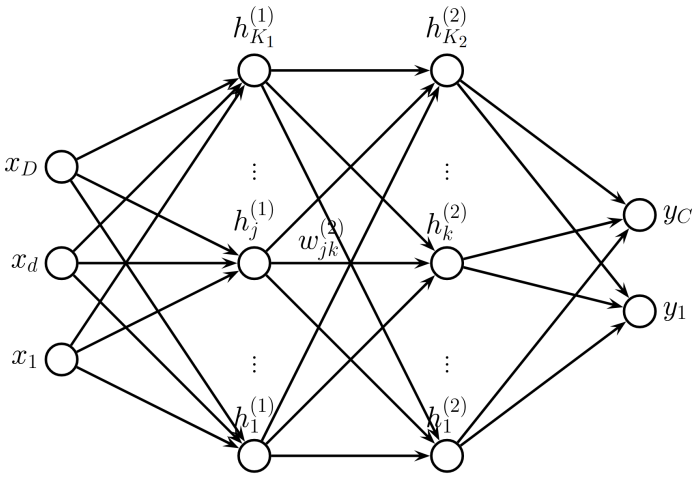


Figure 2.12. Representation of a feedforward neural network. Taken from Murphy 2023.

For example, Figure 2.12 illustrates an MLP architecture with an input layer of D units, two hidden layers with K_1 and K_2 units, and an output layer with C units. The activation of the k-th neuron in hidden layer l is given by:

$$h_k^{(l)} = \varphi_l \left(b_k^{(l)} + \sum_{j=1}^{K_{l-1}} w_{jk}^{(l)} h_j^{(l-1)} \right)$$

where $\varphi_l(\cdot)$ is the activation function at layer l, $b_k^{(l)}$ is the bias term for neuron k and $w_{jk}^{(l)}$ is the weight connecting neuron j in layer $l-1$ to neuron k in layer l. By following this principle, MLPs can be adapted for both classification and regression tasks.

Optimization strategies in regression models

Optimization plays a crucial role in training regression models, ensuring they generalize well to unseen data by minimizing a predefined loss function. In linear regression, the learning process often involves solving an optimization problem using Ordinary Least Squares (OLS).

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This solution exists if $\mathbf{X}^\top \mathbf{X}$ is invertible, ensuring a unique minimiser.

In contrast, random forest regression does not rely on explicit function optimization. Instead, trees are built using a greedy splitting strategy, where at each node, the feature and threshold that minimize variance within the resulting subsets are selected. The final model is an ensemble of independently grown trees, reducing variance through bootstrap aggregation (bagging).

In MLP regression, learning occurs through iterative weight updates using backpropagation, which computes gradients of the loss function via the chain rule. These gradients are then used in Stochastic Gradient Descent (SGD) to update the model parameters (Robbins and Monro 1951). SGD iteratively updates the parameters $\boldsymbol{\theta}$ to minimize the loss function $\mathcal{L}(\boldsymbol{\theta})$, typically the MSE for regression:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta}))^2$$

The gradient-based parameter update

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}),$$

which expands to updating each weight matrix:

$$\mathbf{W}_l \leftarrow \mathbf{W}_l - \eta \frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{W}_l}$$

where η is the learning rate, $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_l}$ is the gradient computed via backpropagation and $l = 1, \ldots, L$ is the number of hidden layers. The implementation of these models is summarised in Chapter A.

2.3.4 Interpretable machine learning techniques

Interpretable machine learning aims to make complex predictive models more transparent by providing insights into how they generate predictions. As machine learning models, particularly deep neural networks and ensemble methods, grow in complexity, their decision-making processes often become opaque, making it difficult to understand the contribution of individual features (Linardatos et al. 2020). This lack of interpretability is a critical limitation, especially in fields like computational biology, where understanding feature importance can reveal underlying biological mechanisms. Various interpretability methods, including feature attribution techniques, surrogate models, and rule-based explanations, have been developed to bridge this gap. These approaches enable researchers to analyse model predictions, assess their reliability, and extract meaningful insights. Among these methods, SHAP (Lundberg and Lee 2017) offers a principled approach to quantifying feature importance based on cooperative game theory (Shapley 1953), ensuring fairness and consistency in feature attributions.

Shapley Additive Explanations (SHAP)

Interpreting complex machine learning models often relies on additive feature attribution, where the model's prediction is approximated as a sum of individual feature contributions. Formally, an explanation model g is defined as a linear function :

$$f(\mathbf{x}) = g\left(\mathbf{x}'\right) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

where x' represents a simplified input space $x = h_x(x')$, M is the number of simplified input features, and ϕ_i denotes the contribution of each feature. This framework ensures that the sum of feature attributions approximates the original model's output $f(x)$. SHAP values are inspired by *Shapley values*, originally developed in cooperative game theory to fairly distribute a total payoff among a set of contributors. Given a predictive model f and a feature set $\mathbf{x}$, the SHAP value for feature x_j quantifies its contribution to the prediction $f(\mathbf{x})$ by computing the average marginal effect of adding

x_j across all possible feature subsets:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

where $|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

Axioms and Theorem 1. SHAP values are the unique solution that satisfies three desirable axioms:

- **Local accuracy (additivity):** The explanation model output equals the original model prediction:
- **Missingness:** If a feature is missing (i.e., $x_i' = 0$), then its attribution is zero: $\phi_i = 0$.
- **Consistency:** If a model changes such that a feature's contribution increases (or stays the same) for all subsets, then its SHAP value should not decrease.

Theorem 1 in Lundberg and Lee 2017 states that SHAP values are the only additive feature attribution method that satisfies all three of these properties. For a general audience, this means that SHAP assigns credit to each feature in a way that is provably fair, faithful to the model, and responsive to any structural improvements in how the model uses a feature. It guarantees that the sum of the parts (the individual feature attributions) always matches the whole (the model's prediction), and that features can't be given credit for outcomes they didn't influence.

Kernel SHAP. Kernel SHAP is a model-agnostic estimator that approximates SHAP values by fitting a weighted linear model over z'.

$$\pi_{x'}\left(z'\right) = \frac{(M-1)}{(M \text{ choose } |z'|)\,|z'|\,(M - |z'|)}$$

This kernel ensures that the solution aligns with the classical Shapley value formulation by weighting feature subsets appropriately in a least-squares regression framework. The loss function is given by:

$$\mathcal{L}\left(f, g, \pi_{x'}\right) = \sum_{z' \in Z} \left[f\left(h_x\left(z'\right)\right) - g\left(z'\right) \right]^2 \pi_{x'}\left(z'\right)$$

which penalizes discrepancies between the model's true output and the linear

explanation model while incorporating the derived kernel weights. This kernel weighting ensures that the regression satisfies the three SHAP axioms. Kernel SHAP is based on LIME (Ribeiro et al. 2016) but derives the loss, regularization, and weighting analytically rather than heuristically.

TreeSHAP. TreeSHAP is a specialized algorithm that computes SHAP values exactly in polynomial time for decision tree models (e.g., XGBoost, LightGBM). Instead of sampling subsets, it recursively traverses each tree and computes the expected contribution of each feature conditioned on its appearance along the tree paths (Lundberg et al. 2019).

The algorithm computes:

$$\phi_i = \mathbb{E}_{z' \sim \text{coalitions without } i} \left[f_x(z' \cup \{i\}) - f_x(z') \right]$$

TreeSHAP calculates these expectations analytically by tracking feature dependencies at each split node and propagating path weights and outputs through the tree structure. It avoids sampling and respects feature dependencies implicitly encoded in the tree, making it accurate and scalable for high-dimensional tabular data.

DeepSHAP. DeepSHAP extends SHAP to deep neural networks by combining DeepLIFT's backpropagation rules with SHAP's theoretical foundation (Lundberg et al. 2019; Shrikumar et al. 2019). It approximates SHAP values by decomposing the output of the network $f(x)$ relative to a reference input x':

$$\phi_i \approx (x_i - x_i') \cdot \frac{\partial f}{\partial x_i}\Big|_{x'}$$

DeepSHAP assigns attributions layer by layer by computing multipliers for each neuron, then propagating these multipliers backwards from the output to the input. These multipliers satisfy the "summation-to-delta" property:

$$\sum_{i=1}^{M} \phi_i = f(x) - f(x')$$

While DeepSHAP is only an approximation to the true SHAP values (because it relies on linearity assumptions between layers and independent inputs), it is much faster than Kernel SHAP and allows interpretation of deep models like MLPs and CNNs. In our study, we use KernelSHAP and DeepSHAP to interpret multilayer

perceptron models trained to predict RNA Polymerase II occupancy. In parallel, we compare these results with TreeSHAP outputs derived from gradient-boosted tree models, leveraging their different assumptions about feature independence and internal structure. This multi-model approach improves the robustness of SHAP-based attributions and helps us identify biologically meaningful regulatory features with greater confidence.

Chapter 3

Research questions

Following the foundational discussion on biology, bioinformatics, and the application of machine learning in data-driven research (Chapter 2), I review recent progress and unresolved gaps in the research.

In the following sections, I identify the resources required for data-driven research and frame the research questions.

3.1 Molecular roles for MeCP2

MeCP2 was originally identified as a methyl-CpG-binding protein through high-affinity pulldown assays that detected proteins binding to modified CpG dinucleotides (Lewis et al. 1992). Subsequent research revealed that MeCP2 missense mutations cause RTT, a severe X-linked neurological disorder (Amir et al. 1999). Due to X-chromosome inactivation, heterozygous females experience a mosaic effect where half of their cells express the mutated protein, leading to disease manifestation. In contrast, hemizygous males do not survive, as they lack a functional copy of MeCP2 (Kankirawatana et al. 2006).

At the molecular level, MeCP2 functions by recruiting the HDAC-containing NCoR co-repressor complex to chromatin through its methylated DNA-binding capability (Ho et al. 2008; Nan et al. 1998). Its mechanism of action involves a chromatin-binding domain and an effector domain that facilitates co-repressor recruitment as introduced in Section 2.1.8. Given its clinical significance, MeCP2 has been extensively studied, leading to a wide array of proposed molecular roles (Lyst and Bird 2015; Tillotson and Bird 2020). These range from binding both methylated and hydroxy-methylated DNA (Ho et al. 2008; Sperlazza et al. 2017), activating and repressing transcription (Chahrour et al. 2008), and influencing alternative splicing through interactions with RNA-binding proteins (Young et al. 2005) to binding tandem repeats (Ibrahim et al. 2021). More recently, MeCP2 has been suggested to bind unmethylated enhancers to repress transcription (Mishra et al. 2025).

Among these proposed roles, the most robustly supported function of MeCP2 is its ability to bind methylated DNA and recruit the NCoR co-repressor complex to regulate transcription. This model is strongly validated by disease-associated mutations that disrupt either of these two functions, leading to phenotypes in mouse models that closely mirror Rett syndrome (Lyst et al. 2013; Lyst and Bird 2015). Additionally,

conditional removal of MeCP2 from all tissues except the brain produces no symptoms in mice, strongly indicating that MeCP2's primary role is in neuronal tissues (Ross et al. 2016). Consequently, studies performed in non-neuronal cell types should be interpreted with caution. Although initially characterized as a methyl-CpG binder, MeCP2 has been found to interact with non-CG methylation (Chen et al. 2015; Gabel et al. 2015), which accumulates in mammalian neurons over time (He and Ecker 2015; Lister et al. 2013). MeCP2 leverages this non-CG methylation as a regulatory signal to fine-tune the neuronal transcriptome (Kinde et al. 2016; Lagger et al. 2017). Several independent studies support this model, showing that modified version of MeCP2 which is unable to bind non-CG methylation develop symptoms similar to Rett syndrome and exhibit overlapping transcriptional changes (Tillotson et al. 2021). Furthermore, mutations in DNMT3A, the enzyme responsible for depositing non-CG methylation, also produce Rett-like symptoms and transcriptomic alterations in mice (Lavery et al. 2020), reinforcing the role of non-CG methylation in MeCP2 function.

Using *in vitro* data, multiple research groups have demonstrated that MeCP2 binds preferentially to methylated cytosines, particularly mCG and mCAC (Gabel et al. 2015; Ho et al. 2008; Lagger et al. 2017). Given its exceptionally high abundance in neuronal cells–comparable to histone proteins–it has been proposed that the number of MeCP2 molecules may exceed its available binding sites (Skene et al. 2010). When fluorescently labelled MeCP2 is visualized under a microscope, it is predominantly enriched in heterochromatic foci, which are repetitive DNA regions rich in DNA methylation (Thakur et al. 2021). With the advent of high-throughput sequencing, *in vivo* assays have been employed to profile MeCP2 occupancy. However, these methods do not reveal strong enrichment at specific loci but rather show a widespread distribution across the genome. When integrated with epigenomic profiling, MeCP2's global binding pattern emerges, aligning with the density of methylated cytosines over broad genomic regions. The first evidence supporting MeCP2's role in transcriptional regulation linked to methylation came from a study suggesting that MeCP2 represses long genes (Gabel et al. 2015). This conclusion, especially the "long gene effect" has since been challenged (Raman et al. 2018). Since MeCP2 binds proportionally to the number of methylated sites across the genome, it has also been shown that its occupancy is influenced by the total number of methylated sites within a gene body (Kinde et al. 2016). This suggests that the observed correlation between MeCP2-mediated repression and long genes may instead be a by-product of multicollinearity among gene length, the number of methylated sites, MeCP2 occupancy, and transcriptional repression. Further insights came from profiling histone acetylation alongside MeCP2 occupancy and transcriptomic changes, revealing

that alterations in histone acetylation and transcriptional activity are also correlated—pointing to the role of NCoR co-repressor activity in MeCP2-mediated regulation (Boxer et al. 2020).

Despite strong evidence for MeCP2's primary role in methylation-dependent transcriptional regulation via NCoR recruitment, alternative functions have been proposed. One such hypothesis suggests that MeCP2 binds RNA-binding proteins to influence alternative splicing (Young et al. 2005). A related idea posits that DNA methylation levels themselves impact splicing regulation, and since MeCP2 binds methylated DNA, it was proposed as a mediator of this process (Lev Maor et al. 2015). However, this phenomenon has been primarily studied in non-neuronal tissues, limiting its relevance to Rett syndrome pathophysiology (Wong et al. 2017). To establish its significance, it is essential to investigate MeCP2's role in splicing regulation specifically in neuronal cells. I address this question in Section 4.1.2 using high-throughput data and applied machine learning techniques.

Although structural studies have confirmed that MeCP2 binds methylated CpG dinucleotides (Ho et al. 2008), and structural modelling suggests that its methylated DNA-binding domain can also recognize non-CG methylation (Lagger et al. 2017), additional evidence indicates that MeCP2 interacts with short tandem repeats (Ibrahim et al. 2021). *In vivo* data suggest that short tandem repeat interaction occurs in neuronal cells, but its functional relevance remains unclear. It is possible that MeCP2 regulates transcription through tandem repeat binding, but this has not been conclusively demonstrated. Using high-throughput sequencing data sets from multiple assay types profiled in neuronal tissues, I address whether CA repeats are primary targets of MeCP2 in Section 4.2.2.

3.2 Short DNA motifs as regulators of transcription

CGIs are short stretches of DNA that serve as hallmarks for promoters in nearly all protein-coding genes in vertebrates (Bird 1986). These regions are characterized by a high abundance of short sequence motifs, predominantly CG dinucleotides. Despite the fact that most CG dinucleotides in vertebrate genomes are methylated, CGIs remain largely unmethylated (Deaton and Bird 2011). Given their widespread presence at gene promoters, it is reasonable to assume that proteins must be able to recognize these regions through sequence-specific binding. Decades of research have validated this assumption, leading to the identification of numerous proteins containing CXXC domains that specifically bind to unmethylated CpG sequences (Blackledge et al. 2010; Denissov et al. 2014; Farcas et al. 2012; Shin Voo et al. 2000; Tate et al. 2010; Thomson

et al. 2010; Wu et al. 2013).

Building on this concept, an alternative hypothesis shifts the focus toward AT-rich sequences. Several proteins are known to recognize short AT motifs, including "AT-hook" proteins. Although the DNA-binding domains of these proteins have been biochemically characterized and their interactions with chromatin-modifying enzymes documented, a complete mechanistic understanding of their function remains elusive (Quante and Bird 2016). One prominent group of AT-rich motif readers is the Spalt-like transcription factor (SALL) family (Yamashita et al. 2007). SALL1 and SALL4 have been implicated in human disease (Kohlhase et al. 1998; Sakaki-Yumoto et al. 2006) and cancers (Gao et al. 2011). These proteins contain a conserved amino acid motif that facilitates the recruitment of the NuRD complex (Lauberth and Rauchman 2006), as introduced in Section 2.1.8. In the Bird Lab, we set out to test the hypothesis that AT motifs or overall DNA base composition can serve as signals to fine-tune transcription. Leveraging high-throughput sequencing data from multiple assays, I have applied regression-based methods to quantify the relationship between gene body AT-content and transcriptional regulation by SALL4 in Section 5.2.

3.3 Regulation of transcription by chromatin-associated proteins

Transcription is regulated by a diverse set of proteins. Here, I use "chromatin-associated" proteins as a broad term encompassing DNA-binding proteins, histone-modifying enzymes, and other proteins that either interact with these factors or function as part of the same complex. The "one gene, one protein" hypothesis (Beadle and Tatum 1941), while historically influential, is an oversimplification. Decades of research have only reinforced the understanding that proteins function in coordination, often within complexes (Alberts 1998). Advancements in technology have enabled the systematic identification of these complexes through various methods (Barrios-Rodiles et al. 2005; Gingras et al. 2007; Jain et al. 2011; Nepusz et al. 2012; Puig et al. 2001). Techniques like ChIP-seq and CUT&RUN, introduced in Section 2.2.2, are commonly used to map chromatin occupancy. However, due to the high cost of antibodies and large-scale "omics" assays, profiling is often limited to a single key subunit—typically a DNA-binding protein or a catalytic enzyme. While this approach provides valuable insights, it assumes co-occupancy of the remaining subunits, an assumption complicated by the fact that components of one complex are often shared across multiple complexes or interact with distinct regulatory assemblies.

Genome engineering tools, discussed in Section 2.1.10, now allow rapid protein

degradation, enabling the investigation of direct transcriptional effects. However, interpreting these results at the level of entire complexes remains challenging due to functional redundancies and the presence of multiple enzymatic activities within a single complex. A prime example is the NuRD complex, which exhibits both chromatin remodelling and histone deacetylase activity (Section 2.1.8). Despite these challenges, researchers have developed innovative strategies to dissect complex functions, including disrupting protein-protein interactions through mutagenesis or systematically profiling multiple subunits to identify shared genomic signatures. Traditional genomic analysis has provided a framework to assess whether multiple proteins co-localize at the same genomic regions, offering clues about their collective function. High-throughput sequencing and machine learning have further expanded these capabilities, allowing for the prediction and imputation of missing signals (Schreiber et al. 2020), leading to a more comprehensive understanding of transcriptional regulation.

In this study, I evaluate the effectiveness of XAI approaches in uncovering regulatory mechanisms by leveraging rapid perturbation datasets generated using Auxin-Inducible Degron (AID) and degradation tag (dTAG) systems. I aim to predict RNA Polymerase II occupancy based on the co-occupancy of chromatin-associated proteins and epigenomic marks, using SHAP (Lundberg and Lee 2017) to quantify the contribution of each factor to transcriptional regulation. By systematically comparing direct and non-direct targets following factor depletion, I assess the reliability of SHAP-derived feature importance scores in inferring regulatory mechanisms of transcription in Section 6.2.

3.3.1 The role of PRC1 and PRC2 in transcription

Polycomb complexes, PRC1 and PRC2, contain catalytic cores that associate with auxiliary proteins to form multiple specialized assemblies (Section 2.1.6). Although these complexes include context-specific regulators, I refer to them collectively as PRC1 and PRC2. For an extensive review of their diversity, refer to Blackledge and Klose 2021. Polycomb complexes are recruited to target sites via sequence-specific DNA-binding factors, primarily through CpG island recognition mediated by CXXC domain-containing proteins. Alternatively, long non-coding RNAs, such as XIST, facilitate Polycomb recruitment, as seen in X-chromosome inactivation (Almeida et al. 2017). While XIST-driven Polycomb targeting results in widespread chromatin repression across the X chromosome, Polycomb enrichment on autosomes is more localized, typically restricted to gene regulatory elements.

Early models suggested that PRC1 represses transcription by binding to H3K27me3 and compacting chromatin, thereby restricting access to gene regulatory elements.

This aligns with observations that PRC1 disruption leads to widespread transcriptional dysregulation. Feedback mechanisms between PRC1 and PRC2, as well as their antagonistic relationship with the H3K4me3-depositing Trithorax system, suggest that CGIs help establish bistable chromatin states at gene promoters (Sneppen and Ringrose 2019). The concept of "bivalency"—where genes simultaneously carry activating (H3K4me3) and repressive (H3K27me3) marks—has been widely explored (Voigt et al. 2013), though its role remains debated. Recent analyses (Shah et al. 2021) question prior ChIP-based interpretations, showing that both marks are present at many loci beyond developmental genes. While mathematical models of chromatin bistability (Sneppen and Ringrose 2019) provide a conceptual framework for bivalency, further validation in mammalian stem cells is required (Macrae et al. 2023).

To date, PRC1 and PRC2 transcriptional effects have been primarily studied in mutant cell lines with genetic knockouts or mutations. However, this approach complicates the distinction between direct and indirect effects (Section 2.1.10). To address this, recent studies have employed rapid protein degradation techniques to acutely deplete RING1B (PRC1 catalytic subunit) and SUZ12 (PRC2 catalytic subunit), allowing the identification of direct transcriptional targets (Dobrinić et al. 2021). Building on these approaches, I aim to use deep learning models to predict Pol–II occupancy from PRC1 and PRC2 binding profiles. Furthermore, I will apply interpretable machine learning techniques to quantify the contribution of PRC1 and PRC2 to transcription at their direct targets, as detailed in Chapter 6.

3.3.2 The role of INTS11 in transcription

Integrator Complex Subunit 11 (INTS11) is a catalytic subunit of the Integrator complex, originally identified through its association with the Pol–II C-terminal domain (CTD) (Baillat et al. 2005). Early studies suggested that Integrator functioned specifically at snRNA genes, but more recent findings have expanded its role to nearly all RNA species (Wagner et al. 2023). This raises fundamental questions about the specificity of Integrator activity and its broader impact on transcription. Long-term depletion of Integrator in mammalian cells leads to both upregulation and downregulation of hundreds of protein-coding genes, suggesting that its function is context-dependent (Elrod et al. 2019). Recent degron-mediated depletion of INTS11 in mouse embryonic stem cells (mESCs) has provided further insight into its role, revealing that acute loss of INTS11 results in a global increase in Pol–II release from promoter regions into gene bodies (Stein et al. 2022). This suggests that Integrator broadly regulates the transition of Pol–II into productive elongation.

A recent study (Wang et al. 2023) proposed that H3K4me3 is essential for the

recruitment of INTS11, linking Integrator activity to the SET1/COMPASS complex, which deposits H3K4me3. To explore this relationship, the study profiled the occupancy of SET1/COMPASS components and INTS11 and identified direct targets using degron-mediated depletion. Acute depletion of SET1/COMPASS components resulted in the downregulation of thousands of genes across multiple RNA profiling techniques (Hughes et al. 2023; Wang et al. 2023), whereas INTS11 depletion led to a mixed response, with both upregulated and downregulated transcripts (Stein et al. 2022; Wang et al. 2023). To dissect these interactions further, I aim to develop a deep learning model to predict Pol–II occupancy and apply interpretable machine learning techniques to quantify the contributions of SET1/COMPASS components and INTS11 to transcriptional regulation.

3.3.3 The role of ZC3H4 in transcription

The human genome is pervasively transcribed by Pol–II, yet much of the transcriptome consists of unstable non-coding RNA (ncRNA) that undergo rapid degradation. Short ncRNAs, including promoter upstream transcripts (PROMPT) and enhancer RNA (eRNA), are terminated early by the Integrator complex (Baillat et al. 2005; Lai et al. 2015; Lykke-Andersen et al. 2021), while protein-coding genes require polyadenylation site (PAS) cleavage for proper termination (Eaton and West 2020). Zinc Finger CCCH-Type Containing 4 (ZC3H4), in complex with WD repeat domain 82 (WDR82), forms the "restrictor" complex (Estell et al. 2023), which plays a crucial role in ncRNA attenuation by preventing excessive elongation at CGIs promoters, a function antagonized by SETD1A/B (Hughes et al. 2023). ZC3H4 and WDR82 help control non-coding RNA transcription in metazoans, with Drosophila Sus acting as the equivalent of ZC3H4 to prevent excessive RNA production. In yeast, a similar protein, Swd2, works with other complexes to stop non-coding RNAs with no sequence conservation with ZC3H4. Overall, yeast and human ncRNA termination pathways share features and functionally analogous components but have little sequence similarity. For a complete overview of ZC3H4 and the Restrictor complex, please refer to Estell and West 2025.

The interaction domain of WDR82 binds to both ZC3H4 and SET1A, but in a mutually exclusive manner, preventing simultaneous interactions with both proteins (Estell et al. 2023). ZC3H4 primarily functions to restrict unproductive transcription, while SET1A plays a key role in transcriptional activation by depositing H3K4me3 at gene promoters. These opposing functions suggest a regulatory balance between ZC3H4-mediated repression and SET1A-mediated activation. Recent studies have shown that the SET1A-WDR82 interaction antagonizes ZC3H4 activity at

protein-coding genes, raising the question of whether these proteins co-occupy the same genomic loci and how their interplay influences transcriptional regulation (Hughes et al. 2023). To investigate this, I analysed chromatin occupancy data in mESCs following rapid degradation of ZC3H4, SET1A, or both factors in combination. By comparing Pol–II occupancy changes across different perturbation conditions, I aimed to quantify the relative contributions of SET1A and ZC3H4 in regulating transcriptional output.

Chapter 4

Reassessing the molecular roles of MeCP2

As outlined in Chapter 3, reproducible evidence for MeCP2's role as a DNA-methylation dependent transcriptional repressor is documented. However, several alternate roles for MeCP2 have also been hypothesised. This chapter attempts to reinvestigate two such hypotheses: MeCP2 as a regulator of alternative splicing, and CA repeats are key mediators of MeCP2 function. Experiments performed to address these hypothesis and their results are presented and attached as is from peer-reviewed manuscripts.

4.1 MeCP2's role in influencing alternative splicing

4.1.1 Introduction

Absence of neuronal MeCP2 is the primary contributor to Rett syndrome (RTT) phenotypes, as the absence of MeCP2 in peripheral tissues does not result in significant phenotypic defects in mice (Ross et al. 2016). Inefficient splicing phenomenon was observed during granulocyte differentiation (Wong et al. 2013) and its mechanistic link with DNA methylation and MeCP2 (Wong et al. 2017) was proposed, but the studies did not investigate the impact of neuronal MeCP2 and DNA methylation on alternative splicing. We argue that if MeCP2 functions as a global regulator of alternative splicing contributing to the Rett Syndrome phenotype, comparable splicing inefficiencies should be detectable in neuronal cells. Therefore, we quantitatively investigate the role of MeCP2 and DNA-methylation in alternative splicing across relevant datasets obtained from neuronal cells. Using modern machine learning approaches alongside traditional bioinformatics analyses, my manuscript (Chhatbar et al. 2020)[1] attempts to quantify linear and non-linear associations between MeCP2, DNA-methylation and alternative splicing.

4.1.2 Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing

[1] This manuscript will appear as is where the experiments have been performed by myself, and the material is collaboratively written by myself, Adrian Bird and Guido Sanguinetti.

PLOS GENETICS

RESEARCH ARTICLE

Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing

Kashyap Chhatbar[1,2], **Justyna Cholewa-Waclaw**[2], **Ruth Shah**[2], **Adrian Bird**[2], **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

* gsanguin@sissa.it

Check for updates

OPEN ACCESS

Citation: Chhatbar K, Cholewa-Waclaw J, Shah R, Bird A, Sanguinetti G (2020) Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing. PLoS Genet 16(10): e1009087. https://doi.org/10.1371/journal.pgen.1009087

Editor: Dirk Schübeler, Friedrich Miescher Institute for Biomedical Research, SWITZERLAND

Received: May 28, 2020

Accepted: August 28, 2020

Published: October 13, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009087

Copyright: © 2020 Chhatbar et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The article is based on publicly available data and accession codes are within the manuscript.

Abstract

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.

Author summary

Rett Syndrome (RTT) is a devastating neurological disorder affecting approximately 1 in 10,000 female births. Most cases of RTT are caused by mutations in the gene identified as methyl-CG binding protein 2 (*MECP2*) which is an epigenetic reader of DNA methylation. Although the primary function of MeCP2 is to recruit NCoR to methylated sites in the genome, the downstream effect on gene expression is subtle and multiple additional functions have been proposed. Here we focus on the influence of MeCP2 on one of these: alternative splicing to generate different messenger RNAs from a single primary transcript. Using machine learning approaches, we show that neither MeCP2 nor DNA methylation influence alternative splicing. Our results emphasize the importance of multivariate quantitative analyses and they challenge the over-interpretation of causal relationships based on high-throughput data sets.

PLOS GENETICS

Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing

Funding: Kashyap Chhatbar was supported by a scholarship from College of Science and Engineering, University of Edinburgh. This work was supported by Wellcome Centre grant (091580/Z/10/Z), a Wellcome Investigator Award (107930/Z/15/Z), the Rett Syndrome Research Trust and a European Research Council Advanced grant (EC 694295 Gen-Epix) to Adrian Bird. Adrian Bird is a member of the Simons Initiative for the Developing Brain at the University of Edinburgh. Ruth Shah was supported by a Wellcome Trust 4-y PhD studentship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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, our previous work [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. Additionally, Boxer et al. [17] comprehensively captured transcriptomes and methylomes from wildtype (*WT*) and MeCP2 (*KO*) mouse brain tissues. To investigate the latter, we examined the *Dnmt1/3a/3b* triple-knockout mouse

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

Accession	Source	Total RNA-seq Coverage	Reps	Type
GSE125660 [9]	MeCP2 *WT* Neurons	218.2×10^6 pairs	4	BS-seq, RNA-seq, ChIP-seq
	MeCP2 *KO* Neurons	214.2×10^6 pairs	4	RNA-seq, ChIP-seq
	MeCP2 *OE4x* Neurons	217.5×10^6 pairs	4	
	MeCP2 *OE11x* Neurons	220.4×10^6 pairs	4	
GSE128186 [17]	*WT* brain tissue	713.9×10^6 reads	10	BS-seq, RNA-seq
	MeCP2 *KO* brain tissue	796.5×10^6 reads	10	
GSE64910 [18]	*WT* mESCs	49.1×10^6 reads	2	BS-seq, RNA-seq
	DNMT-TKO mESCs	50.3×10^6 reads	2	
GSE67867 [19]	*WT* mESCs	357.3×10^6 reads	3	RNA-seq
	DNMT-TKO mESCs	364.8×10^6 reads	3	
GSE103214 [20]	*Pv* Neurons (1 week)	37.7×10^6 reads	2	BS-seq, RNA-seq
	Pv Neurons (3 weeks)	27.6×10^6 reads	1	
	Pv Neurons (8 weeks)	97.5×10^6 reads	2	
	Vip Neurons (1 week)	17.9×10^6 reads	1	
	Vip Neurons (3 weeks)	57.9×10^6 reads	2	
	Vip Neurons (8 weeks)	45.3×10^6 reads	2	
GSE48307 [23] GSE85517 [14]	Promyelocytes	63.5×10^6 reads	1	BS-seq, RNA-seq, ChIP-seq
	Granulocytes	59.7×10^6 reads	1	

https://doi.org/10.1371/journal.pgen.1009087.t001

embryonic stem cells (*DNMT-TKO* mESCs) that lack DNA methyltransferase activity [18, 19]. Yearim et al. performed splicing analyses on a relatively low-coverage data set and suggested that DNA methylation may influence splicing of alternate exons [18]. Therefore, we sought to revisit this conclusion using a more recent, higher coverage *DNMT-TKO* dataset [19] to investigate the role of DNA methylation in influencing alternative splicing independent of MeCP2. Finally, it is also possible that splicing regulation is due to a combination of changes in MeCP2 levels and DNA methylation. In early postnatal brain, the level of genomic mCA, a major determinant of MeCP2 binding, concurrently increases with the level of MeCP2 [5, 20, 21]. Stroud et al. [20] profiled the transcriptomes and methylomes of differentiating mouse neurons in the first few weeks after birth. Therefore, developing neurons can be considered an ideal scenario where one can investigate the combined effects of MeCP2 and DNA methylation. For completeness, we also re-analysed the non-neuronal data sets from promyelocytes and granulocytes [14] which first led to suggest the potential link between MeCP2, DNA methylation and alternative splicing. MeCP2 RNA expression levels show predominant expression in neurons and brain tissue while mESCs and granulocyte express significantly lower levels (S1 Fig). All of these data sets provide a good level of replication and high sequencing depth, with the exception of the early developmental time points of [20], where experimental constraints limited the achievable depth.

Adequate sequencing coverage is important for robust alternative splicing analysis, and some of the data sets we employ, particularly the early developmental times in the Stroud et al. data [20], have relatively low coverage. To address this, we use a recently proposed splicing quantification strategy, BRIE (Bayesian Regression for Isoform Estimation [22]), which uses Bayesian sequence-derived, informative priors to provide robust estimates of splicing ratios even at very low coverage levels such as encountered in single cell RNA-seq data.

Summary statistics for all the data sets used, including sequencing depth and replication level, are provided in Table 1.

PLOS GENETICS Quantitative analysis questions the role of MeCP2 as a global regulator of 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) (S2A and S2B Fig). We observe lower number of differential splicing events in *KO* mouse brain (57) compared to changes in gene expression (2943) (S2C and S2D Fig). This suggests that changing the levels of MeCP2 (*in vitro* or *in vivo*) 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 lower 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 (S2E and S2F Fig). 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) (S2E and S2F Fig). 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 [20, 21] and the dynamic nature of alternative splicing [24] 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) (S2G and S2H Fig). 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 [18], the number of differentially regulated genes (102) was only marginally higher than number of differential splicing events (71) (S2I and S2J Fig). Differential gene expression analysis was not possible on the RNA-seq data set of Promyelocytes and Granulocytes [14, 23] due to the absence of replicates. The lower number of differential splicing events compared to differentially expressed genes is consistent at different significance thresholds (S2 Fig) in all comparisons except in developing neurons. To investigate whether differential splicing events are directly associated with context specific DNA methylation or MeCP2 binding, we performed correlation analysis on significanctly altered events (S2B, S2D, S2F, S2H and S2J Fig). We found that there is no apparent linear relationship between changes in splicing and changes in DNA methylation or MeCP2 binding (S1 Table) except a modest one for *Pv* neurons (highlighted in S1 Table). 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 multivariate 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 [25, 26]. The splicing ratios were calculated using several independent methods, and our analysis indicated a strong correlation between results obtained with different methods (S3A and S3E Fig). Methods that directly model splicing ratios such as Miso [27] and BRIE [22] showed particularly strong agreement (S3A Fig). 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 [25, 26] could be confirmed for all the data sets described in Table 1 We find that both models explain 50%-65% of the variation (Fig 1A, 1B, 1D, 1E and 1F 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 (Fig 1A, 1B, 1D, 1E and 1F right panels). We note that predictive models achieved lower accuracy in *Vip* neurons and *Pv* neurons during early mouse brain development (Fig 1C), possibly due to the highly dynamic nature of splicing in early development [24]. 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 (Fig 1) [25, 26], our hypothesis is that any role of MeCP2 in regulating splicing should be reflected in improved predictions of differential exon inclusion/exclusion. Since global DNA methylation remains constant in neurons expressing multiple levels of MeCP2 (S4A Fig), we assume that DNA methylation at exons and introns also remain constant. To allow for the possibility that 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 (*in vitro*) 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 (Fig 2A left panel). In some instances, the model fit is extremely poor, as evidenced by negative levels of splicing variation explained (Fig 2A right panel). To recapitulate a similar situation *in vivo*, we used data sets from [17] which profiled transcriptomes and methylomes of *WT* and MeCP2 *KO* mouse brain tissues at very high sequencing depth. We found that DNA methylation levels, quantified from WGBS-seq, are statistically indistinguishable between *WT* and MeCP2 *KO* mice brains (S4B and S4C Fig): correlations between methylomes of different WT replicates are identical to correlations between methylomes in *WT* and *KO* samples. This further supports our assumption that DNA methylation remains constant in neurons expressing different levels of MeCP2. In mouse brain data sets, when we quantified the explanatory power of MeCP2 by regressing splicing variation against sequence features, the model fit was once again poor, with negligible or even negative levels of splicing variation explained (Fig 2B). Importantly, the 57 differentially spliced genes in this data set did not show any significant level of differential methylation (S1 Table), ruling out a potentially confounding effect of a compensatory local change in methylation. Therefore, multiple levels of MeCP2 (*in vitro* and

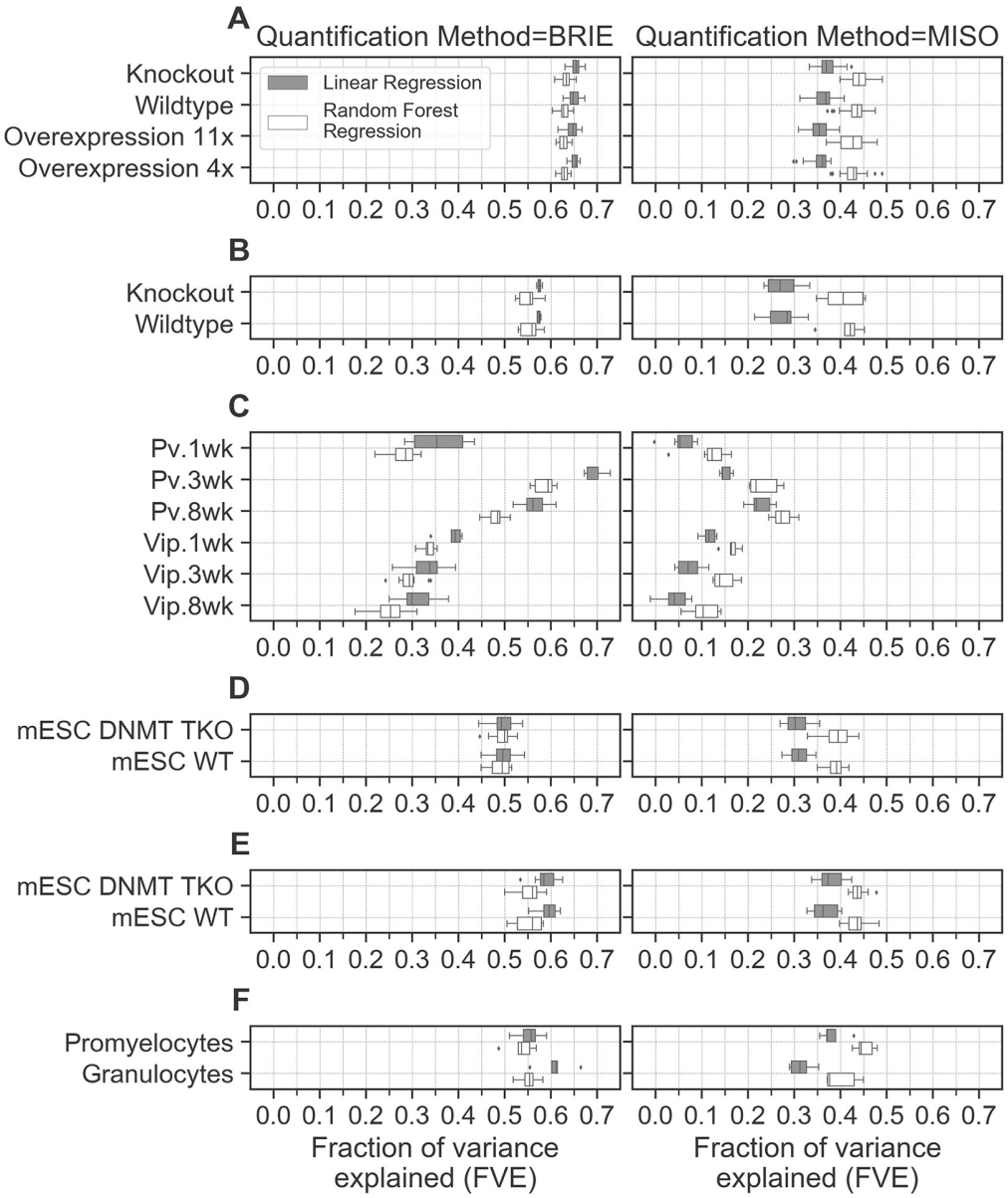


Fig 1. Fraction of variance explained by regressing sequence features against splicing ratios. (A) Cultured human neurons expressing multiple levels of MeCP2 [9]**(B)** Wildtype (*WT*) and MeCP2 Knockout (*KO*) brain tissue [17]**(C)** *Pv* and *Vip* neurons at 1 week, 3 weeks and 8 weeks [20] **(D)** *DNMT-TKO* and *WT* mESCs [18]**(E)** *DNMT-TKO* and *WT* mESCs [19] **(F)** Promyelocytes and Granulocytes [14, 23]. See Table 1 for detailed information about data sets.

https://doi.org/10.1371/journal.pgen.1009087.g001

in vivo) combined with 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,

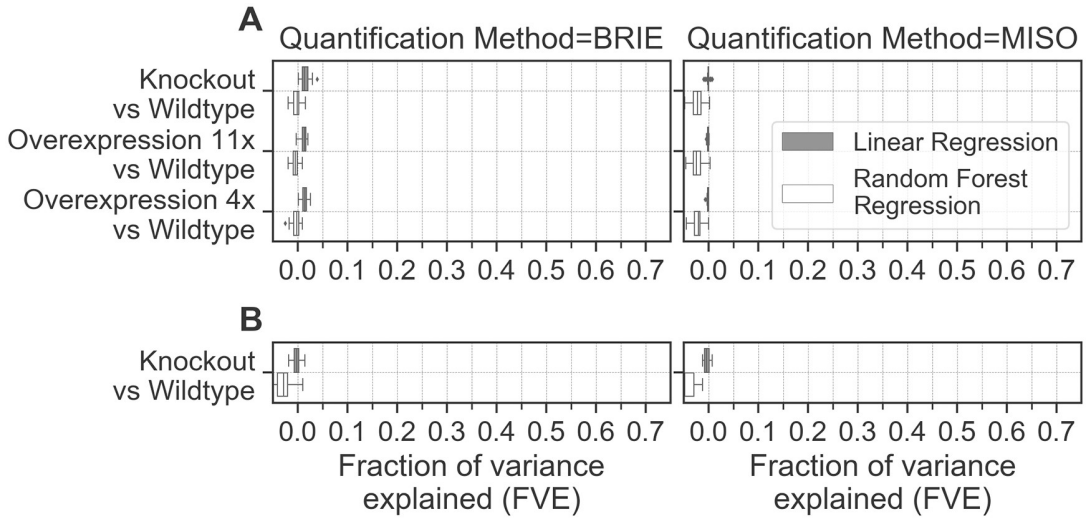


Fig 2. Explanatory power of varying MeCP2 levels (*in vitro* and *in vivo*). (A) Fraction of variance explained by regressing splicing variation against sequence features combined with MeCP2 binding in cultured human neurons [9]. **(B)** Fraction of variance explained by regressing splicing variation against sequence features combined with changes in DNA methylation in *WT* and MeCP2 *KO* mouse brain [17].

https://doi.org/10.1371/journal.pgen.1009087.g002

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 (Fig A in S1 Appendix) in developing mouse neurons [20]. 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% (Fig 3A left panel). The analysis fails to capture any relationship (linear/non-linear) when splicing variation is regressed against DNA methylation alone (S5A Fig). 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 independently of MeCP2. 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 [18, 19]. As these cells also express low levels of MeCP2 (S1 Fig), this analysis is important to rule out any confounding effects on our conclusions possibly caused by the high correlation between MeCP2 binding and DNA methylation in neuronal cells. 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 (Fig 3C left panel). In a separate *DNMT-TKO* mESCs data set, the accuracy was lower (0-5%) (Fig 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 only one where methylation levels combined with sequence features do appear to explain a small fraction of variance in splicing

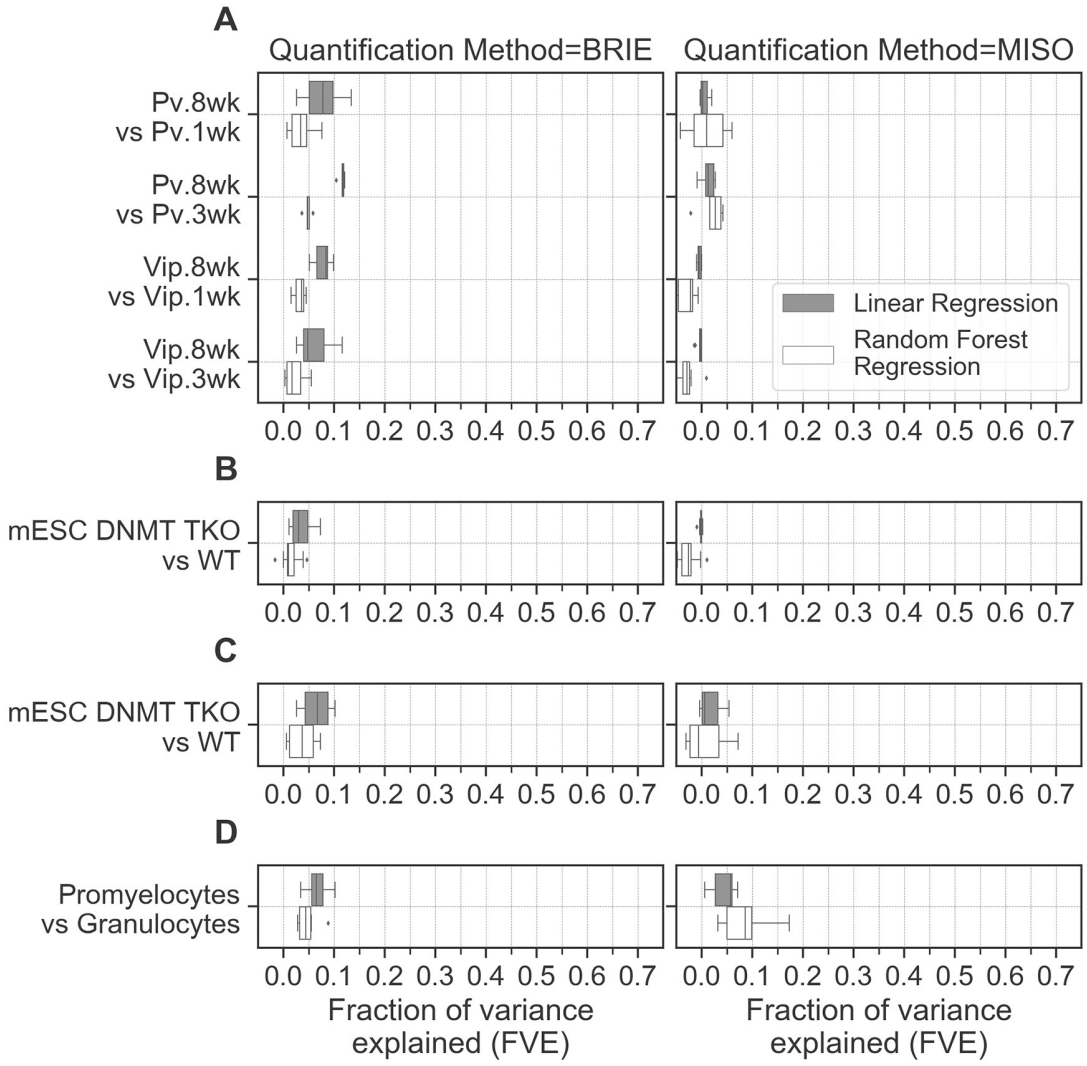


Fig 3. Fraction of variance explained by regressing splicing variation against sequence features along with DNA methylation features. (A) *Pv* and *Vip* neurons at 1 week, 3 weeks and 8 weeks [20]**(B)** *DNMT-TKO* and *WT* mESCs [19]**(C)** *DNMT-TKO* and *WT* mESCs [18]**(D)** Promyelocytes and Granulocytes [14, 23].

https://doi.org/10.1371/journal.pgen.1009087.g003

variation (5-10%) (Fig 3D). 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, Fig 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 (S3B Fig), 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. It should however be pointed out that our analysis of splicing variation of *WT* and MeCP2 *KO* mouse brains highlighted that changes in splicing were not associated with any significant changes in DNA methylation at the same loci, suggesting that even a gene-specific role of MeCP2/DNA methylation in regulating splicing is likely to be complex and/ or indirect. 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 specifically in non-brain tissues has no obvious phenotypic consequences, indicating that the protein is dispensable outside the central nervous system [28].

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.

Materials and methods

Sequencing data sets

Table 1 details the data sets used in the analysis. These include RNA sequencing (RNA-seq) and Bisulfite 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. Information about genome assemblies, annotation, software versions and command-line arguments are detailed in S1 Appendix.

Quantification of splicing ratios

RNA-Seq data sets are downloaded and sequencing reads are mapped to the reference genome using HISAT2 [29] aligner. Splicing ratios are quantified using Miso [27], BRIE [22] and Cufflinks [30]. We found Cufflinks to perform poorly in simulated data with known Ψ values (S3D Fig) when the sequence reads were single-end (S1 Appendix). The Pearson correlation between Cufflinks and Miso quantification in our data sets reflects this (representative scatter plot S3E Fig). Because two of the data sets that we have considered are sequenced with single-end reads and the resultant suboptimal modelling (S3F Fig), we decided to omit Cufflinks quantifications and regression models from further analyses.

Differential gene expression

Gene counts are extracted using featureCounts [31] from alignment files. Subsequently, differential gene expression is estimated using DESeq2 [32].

Differential splicing

Differential splicing is estimated using Miso [27].

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 [33] 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. A visual representation of the specific regions is described in S1 Appendix.

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` [33] for regions around the alternative exon. A visual representation of the specific regions is described in S1 Appendix.

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 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 [34, 35] which is implemented in the `scikit-learn` [36] package. Information about arguments and parameters are detailed in S1 Appendix.

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 [37], implemented in `scikit-learn` [36] 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 the 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 the 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.

Supporting information

S1 Table. Pearson (*r*) correlation between changes in splicing ratios and changes in DNA methylation for all data sets.
(PDF)

S1 Fig. MeCP2 expression across data sets considered for analysis (refer Table 1) as counts per million quantified from RNA-seq data.
(TIF)

S2 Fig. 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 MeCP2 *KO* mouse brains compared to MeCP2 *WT* [17] **(E)** Number of differentially regulated genes and **(F)** Number of differential splicing events in developing mouse neurons [20] **(G)** Number of differentially regulated genes and **(H)** Number of differential splicing events in *DNMT-TKO* mESCs [19] **(I)** Number of differentially regulated genes and **(J)** Number of differential splicing events in *DNMT-TKO*

mESCs [18].
(TIF)

S3 Fig. 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)** Representative 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.
(TIF)

S4 Fig. DNA methylation levels in neurons expressing multiple levels of MeCP2. (A) Referencing results from our previous work [9]. High Performance Liquid Chromatography (HPLC) quantification of methylated cytosines in neurons expressing wild-type, 4 times and 11 times MeCP2. **(B)** CG and **(C)** CA methylation across introns quantified from bisulfite sequencing of *WT* and MeCP2 *KO* mouse brains (GSE128172 [17]).
(TIF)

S5 Fig. Fraction of variance explained by regressing DNA methylation features against differential splicing ratios in (A) developing mouse neurons [20]**(B)** Promyelocytes and Granulocytes [14, 23].
(TIF)

S1 Appendix. Detailed information about software, software versions and command line arguments used for analysis.
(PDF)

Acknowledgments

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

Author Contributions

Conceptualization: Kashyap Chhatbar, Guido Sanguinetti.

Data curation: Kashyap Chhatbar.

Formal analysis: Kashyap Chhatbar.

Funding acquisition: Adrian Bird, Guido Sanguinetti.

Investigation: Kashyap Chhatbar, Guido Sanguinetti.

Methodology: Kashyap Chhatbar, Justyna Cholewa-Waclaw, Ruth Shah.

Project administration: Guido Sanguinetti.

Resources: Justyna Cholewa-Waclaw, Ruth Shah.

Software: Kashyap Chhatbar.

Supervision: Adrian Bird, Guido Sanguinetti.

Validation: Kashyap Chhatbar, Justyna Cholewa-Waclaw, Ruth Shah.

Visualization: Kashyap Chhatbar.

Writing – original draft: Kashyap Chhatbar, Guido Sanguinetti.

Writing – review & editing: Kashyap Chhatbar, Adrian Bird, Guido Sanguinetti.

References

1. Guy J, Cheval H, Selfridge J, Bird A. The Role of MeCP2 in the Brain. Annual Review of Cell and Developmental Biology. 2011; 27(1):631–652.
2. Tillotson R, Bird A. The Molecular Basis of MeCP2 Function in the Brain. Journal of Molecular Biology. 2020; 432(6):1602–1623. https://doi.org/10.1016/j.jmb.2019.10.004
3. Lavery LA, Zoghbi HY. The distinct methylation landscape of maturing neurons and its role in Rett syndrome pathogenesis. Current Opinion in Neurobiology. 2019; 59:180–188. https://doi.org/10.1016/j.conb.2019.08.001
4. Amir RE, Veyver IBVd, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2 Nature Genetics. 1999; 23(2):185–188. https://doi.org/10.1038/13810
5. Hendrich B, Bird A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins Molecular and Cellular Biology. 1998; 18(11):6538–6547. https://doi.org/10.1128/MCB.18.11.6538
6. Lagger S, Connelly JC, Schweikert G, Webb S, Selfridge J, Ramsahoye BH, et al. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. PLOS Genetics. 2017; 13(5):e1006793. https://doi.org/10.1371/journal.pgen.1006793
7. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global Epigenomic Reconfiguration During Mammalian Brain Development. Science. 2013; 341(6146). https://doi.org/10.1126/science.1237905
8. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Research. 2013; 23(3):555–567. https://doi.org/10.1101/gr.147942.112
9. Cholewa-Waclaw J, Shah R, Webb S, Chhatbar K, Ramsahoye B, Pusch O, et al. Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. Proceedings of the National Academy of Sciences. 2019; 116(30):14995–15000. https://doi.org/10.1073/pnas.1903549116
10. Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015; 522(7554):89–93. https://doi.org/10.1038/nature14319
11. Kinde B, Wu DY, Greenberg ME, Gabel HW. DNA methylation in the gene body influences MeCP2-mediated gene repression. Proceedings of the National Academy of Sciences. 2016; 113(52):15114–15119. https://doi.org/10.1073/pnas.1618737114
12. Lyst MJ, Bird A. Rett syndrome: a complex disorder with simple roots. Nature Reviews Genetics. 2015; 16(5):261–275. https://doi.org/10.1038/nrg3897
13. Young JI, Hong EP, Castle JC, Crespo-Barreto J, Bowman AB, Rose MF, et al. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. Proceedings of the National Academy of Sciences. 2005; 102(49):17551–17558. https://doi.org/10.1073/pnas.0507856102
14. Wong JJL, Gao D, Nguyen TV, Kwok CT, van Geldermalsen M, Middleton R, et al. Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. Nature Communications. 2017; 8(1):15134. https://doi.org/10.1038/ncomms15134
15. Aslanzadeh V, Huang Y, Sanguinetti G, Beggs JD. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. Genome Research. 2018; 28(2):203–213. https://doi.org/10.1101/gr.225615.117
16. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends in Genetics. 2015; 31(5):274–280. https://doi.org/10.1016/j.tig.2015.03.002
17. Boxer LD, Renthal W, Greben AW, Whitwam T, Silberfeld A, Stroud H, et al. MeCP2 Represses the Rate of Transcriptional Initiation of Highly Methylated Long Genes Molecular Cell. 2020; 77(2):294–309.e9. https://doi.org/10.1016/j.molcel.2019.10.032

18. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, et al. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Reports. 2015; 10(7):1122–1134. https://doi.org/10.1016/j.celrep.2015.01.038

19. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1 Nature. 2015; 528(7583):575–579. https://doi.org/10.1038/nature16462

20. Stroud H, Su SC, Hrvatin S, Greben AW, Renthal W, Boxer LD, et al. Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. Cell. 2017; 171(5):1151–1164.e16. https://doi.org/10.1016/j.cell.2017.09.047

21. He Y, Ecker JR. Non-CG Methylation in the Human Genome. Annual Review of Genomics and Human Genetics. 2015; 16(1):55–77. https://doi.org/10.1146/annurev-genom-090413-025437

22. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biology. 2017; 18(1):123. https://doi.org/10.1186/s13059-017-1248-5

23. Wong JL, Ritchie W, Ebner O, Selbach M, Wong JH, Huang Y, et al. Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. Cell. 2013; 154(3):583–595. https://doi.org/10.1016/j.cell.2013.06.052

24. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology. 2017; 18(7):437–451. https://doi.org/10.1038/nrm.2017.27

25. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010; 465(7294):53–59. https://doi.org/10.1038/nature09000

26. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease Science. 2015; 347(6218). https://doi.org/10.1126/science.1254806

27. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 2010; 7(12):1009–1015. https://doi.org/10.1038/nmeth.1528

28. Ross PD, Guy J, Selfridge J, Kamal B, Bahey N, Tanner KE, et al. Exclusive expression of MeCP2 in the nervous system distinguishes between brain and peripheral Rett syndrome-like phenotypes Human Molecular Genetics. 2016; 25(20):4389–4404. https://doi.org/10.1093/hmg/ddw269

29. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019; 37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4

30. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJv, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010; 28(5):511–515. https://doi.org/10.1038/nbt.1621

31. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656

32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15(12):550. https://doi.org/10.1186/s13059-014-0550-8

33. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010; 26(17):2204–2207. https://doi.org/10.1093/bioinformatics/btq351

34. MacKay DJC. Bayesian Interpolation. Neural Computation. 1992; 4(3):415–447. https://doi.org/10.1162/neco.1992.4.3.415

35. Tipping ME. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research. 2001; 1(Jun):211–244.

36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–2830.

37. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

4.1.3 Conclusion

We conclude that splicing regulation is not a primary function of MeCP2, and alternative splicing remains unaffected by developmental changes in DNA methylation levels. The findings in our peer reviewed manuscript highlight the requirement of uniform analysis of multi-modal sequencing experiments to test a given hypothesis.

4.2 The Impact of CA Repeats on MeCP2 Function

4.2.1 Introduction

Multiple classes of satellite DNAs are prone to DNA methylation, especially CA repeats (Papin et al. 2017) and some are involved in distinct cellular functions (Thakur et al. 2021). Neuronal MeCP2 binds non-CG methylated DNA (Chen et al. 2015; Gabel et al. 2015; Guo et al. 2014; Lagger et al. 2017) and MeCP2's inability to bind non-CG methylated DNA primarily contributes to the pathogenesis of Rett Syndrome (Lavery et al. 2020; Tillotson et al. 2021). A functional link between CA repeats, DNA methylation and MeCP2 is proposed to play a role in Rett Syndrome pathology (Ibrahim et al. 2021). In my manuscript (Chhatbar et al. 2022) [2], we disentangle whether atypical CA repeats in mouse genome influence the molecular function of MeCP2.

4.2.2 A critique of the hypothesis that CA repeats are primary targets of neuronal MeCP2

[2]This manuscript will appear as is where the experiments have been performed by myself and John Connelly, and the material is collaboratively written by myself, John Connelly and Adrian Bird.

Published Online: 19 September, 2022 | Supp Info: http://doi.org/10.26508/lsa.202201522
Downloaded from life-science-alliance.org on 7 August, 2024

Research Article

TRANSPARENT PROCESS OPEN ACCESS Life Science Alliance

Check for updates

A critique of the hypothesis that CA repeats are primary targets of neuronal MeCP2

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

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

DOI 10.26508/lsa.202201522 | Received 11 May 2022 | Revised 7 September 2022 | Accepted 7 September 2022 | Published online 19 September 2022

Introduction

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

This scenario is challenged by a recent proposal that arrayed tandem repeats of the dinucleotide CA are critical MeCP2 targets, exceeding in importance both mCG and isolated mCA moieties as mediators of MeCP2 function (Ibrahim et al, 2021). As $[CA]_n$ repeat blocks are relatively frequent, it is argued that their proximity to genes provides high-affinity "landing pads" through which MeCP2-dependent gene regulation is mediated. Mechanistically, it is suggested that MeCP2 binding to occasional mCA or hmCA moieties within $[CA]_n$ repeats can seed cooperative MeCP2 binding across the entire array, including non-methylated CA motifs. Here, we further investigate the relationship between MeCP2 and CA repeats. Our findings do not offer support for the claim that cytosine modification is enriched at $[CA]_n$ arrays or that MeCP2 is preferentially bound at these repeat blocks in brain cell nuclei. Moreover, we find that the effects of MeCP2 deficiency on transcription in various brain regions do not correlate with proximity to $[CA]_n$ repeats but instead strongly correlate with local mCAC frequency.

Results

Absence of enrichment of modified cytosine in CA repeats

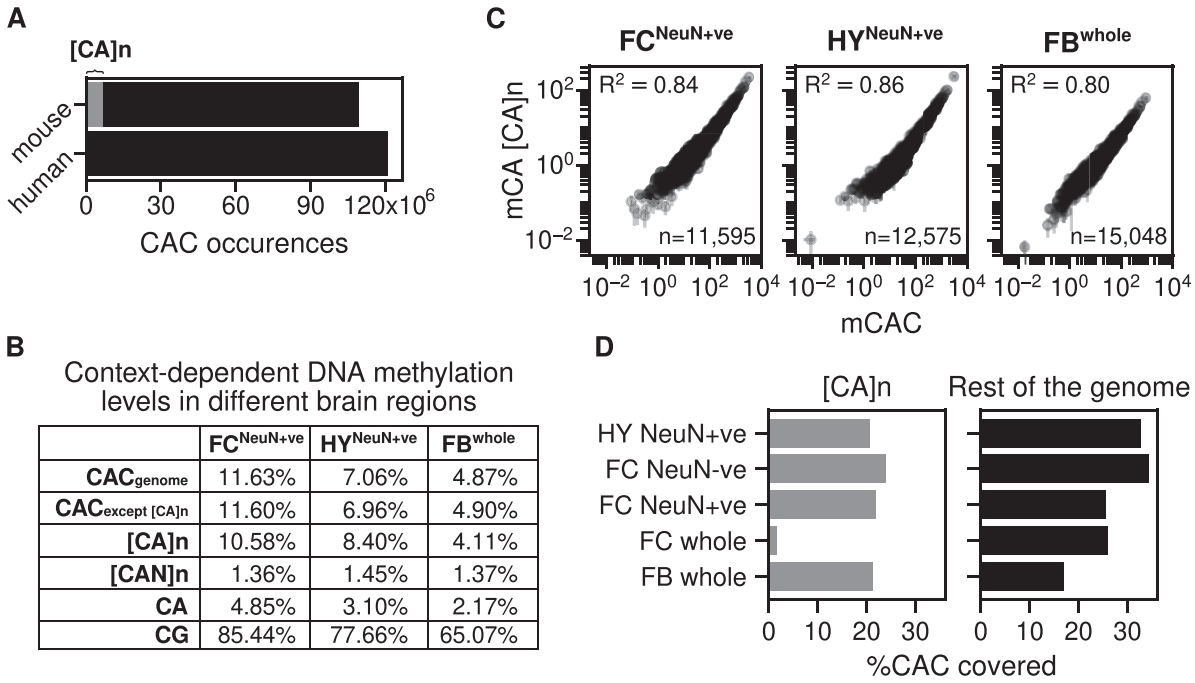
The mouse genome (version mm9) contains ~320,000 $[CA]_n$ arrays (minimum 10 base pairs) of variable length with an average of 25 CA repeats each. It has been reported that CA is more frequently methylated or hydroxymethylated within $[CA]_n$ repeat arrays than elsewhere in the genome (Ibrahim et al, 2021), but this comparison did not take account of the preferential methylation of CAC trinucleotides in neurons (Lagger et al, 2017). Although CAC is necessarily very abundant within $[CA]_n$ repeats, isolated CA motifs elsewhere in the genome may or may not have C in the third position. Recognizing that the trinucleotide sequence CAC is the preferred target of non-CG methylation in neurons and is also a

[1]Wellcome Centre for Cell Biology, University of Edinburgh, The Michael Swann Building, Edinburgh, UK [2]Informatics Forum, School of Informatics, University of Edinburgh, Edinburgh, UK [3]Ludwig Institute for Cancer Research, University of Oxford, Oxford, UK

Correspondence: a.bird@ed.ac.uk
*Kashyap Chhatbar and John Connelly contributed equally to this work.

© 2022 Chhatbar et al.

Life Science Alliance



	FC$^{NeuN+ve}$	HY$^{NeuN+ve}$	FB^{whole}
CAC$_{genome}$	11.63%	7.06%	4.87%
CAC$_{except\ [CA]n}$	11.60%	6.96%	4.90%
[CA]n	10.58%	8.40%	4.11%
[CAN]n	1.36%	1.45%	1.37%
CA	4.85%	3.10%	2.17%
CG	85.44%	77.66%	65.07%

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

target for MeCP2 binding (Lagger et al, 2017; de Mendoza et al, 2021), we determine that ~6% of all CAC motifs in mouse are found within $[CA]_n$ repeats (Fig 1A). Interestingly, humans (version chm13-v1.1) possess a much lower proportion of $[CA]_n$ repeats: ~55,000 $[CA]_n$ arrays amounting to only ~1% of all CAC motifs in the genome (Fig 1A). Using published data for three mouse brain regions, we confirmed that mCA occurs at a lower frequency outside $[CA]_n$, but the frequency of mCAC for each brain region was similar within and outside the repeat arrays (Fig 1B). To determine whether the levels of cytosine modification in $[CA]_n$ arrays match the level of mCAC nearby, we plotted mCAC number per gene against mCA number exclusively within CA repeats (Fig 1C). The results showed a strong correlation, indicating that the local density of mCAC in genes is similar regardless of whether the trinucleotide is isolated or within a CA repeat array. We conclude from these findings that, although $[CA]_n$ arrays are subject to CAC methylation, they are not targeted preferentially compared with the surrounding genome but tend to adopt a level of mCAC that reflects the neighbouring DNA.

This conclusion assumes that the level of cytosine modification in $[CA]_n$ arrays has not been systematically underestimated by bisulfite sequencing because of reduced coverage of repetitive sequences. To test for under-representation in published bisulfite sequence data, we compared the fraction of CAC covered in $[CA]_n$ repeats versus the rest of the genome. The results show little difference for forebrain and reduced coverage of $[CA]_n$ arrays (<2-fold) in NeuN+ve frontal cortex and hypothalamus (Fig 1D). Slightly lower coverage of CA repeats has little effect on estimates of DNA methylation levels as, even when inadequately covered sequences are excluded, the number of CACs that are reliably detected (1,913,430) is more than sufficient to allow accurate determination of their modification status. The exception to this was the whole frontal cortex dataset (280,015 CACs reliably detected), where bisulfite sequence data gave much lower coverage of $[CA]_n$ repeats (Lister et al, 2013). Importantly, only this sparsely covered dataset was analyzed by Ibrahim et al (2021). High bisulfite sequence coverage was obtained with purified cortical NeuN+ve (neuronal) and NeuN−ve (mostly non-neuronal) nuclei from the same study (Lister et al, 2013) (Fig 1D), demonstrating that CA repeats are not intrinsically under-represented in the cortex by this technology. Leaving aside the outlier dataset from whole frontal cortex, the evidence indicates a modest bias by bisulfite sequencing against $[CA]_n$ repeats. Despite this effect, CA repeat coverage is sufficient to strongly support the conclusion that methylation of mCAC is similar between CA repeats and the rest of the genome.

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

We next asked whether MeCP2 is preferentially associated with CA repeat arrays (Ibrahim et al, 2021). MeCP2 ChIP of mouse brain

Life Science Alliance

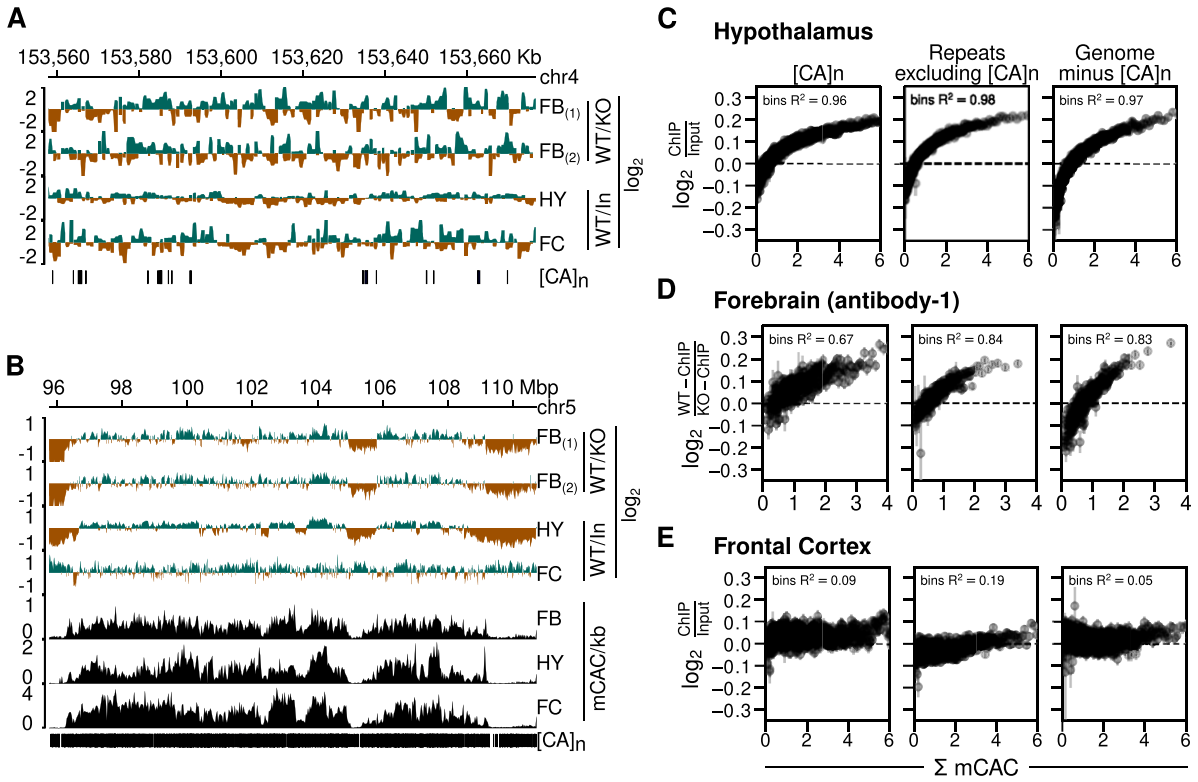


Figure 2. Absence of enhanced MeCP2 binding at CA repeat arrays in brain.
(A) Genome browser screenshots of mouse chr4:153558029-153676791 (version mm9) showing *Mecp2 wild-type* (*WT*) ChIP signal normalized to *Mecp2 null* (*KO*) ChIP signal for the forebrain (FB) using two distinct antibodies (Boxer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

reproducibly reveals relatively uniform genome occupancy with few prominent peaks (Chen et al, 2015; Gabel et al, 2015; Lagger et al, 2017). This has been interpreted to reflect the high frequency throughout the neuronal genome of short MeCP2 target sites, mCG and mCAC (Lagger et al, 2017). In contrast, Ibrahim et al report that MeCP2 ChIP-Seq reads in cultured MEFs are concentrated in prominent peaks coincident with CA repeat clusters (Ibrahim et al, 2021). Given the importance of MeCP2 function and the exceptionally high abundance of hmC and mC in neurons, equivalent peaks at $[CA]_n$ might be expected in the brain. However, published ChIP data do not support this prediction. As an example, the MeCP2-binding profile (normalized to the *KO* or input ChIP profiles) across the same ~120 kb region of the mouse genome that was illustrated for MEFs (Ibrahim et al, 2021) failed to highlight $[CA]_n$ arrays (Fig 2A). In view of uncertainties regarding the initial MEF data (see the Discussion section), our findings question the evidence for preferential localization of MeCP2 to CA repeats.

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

Life Science Alliance

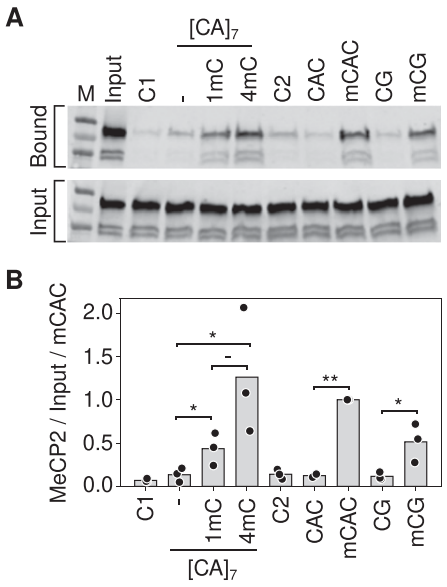


Figure 3. MeCP2 binding to $[CA]_7$ is dependent on cytosine methylation.
(A) Example of a pulldown assay for MeCP2 binding using biotin-tagged double-stranded DNA oligonucleotides incubated with mouse brain nuclear extracts (see the Materials and Methods section). Unrelated probes C1 and C2 contained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

genome categories, but the three DNA sequence types showed closely similar profiles in each brain region. The same conclusion can be drawn from data from the forebrain, which is represented by two high-coverage datasets using independent anti-MeCP2 antibodies (Fig 2D). A fourth dataset derived from mouse frontal cortex (Lister et al, 2013; Gabel et al, 2015) also showed a trio of near-identical ChIP plots, but in this case, the relationship to mCAC frequency was less striking (Fig 2E). Rank correlations derived from the averaged bins across the whole genome excluding CA repeats gave R^2 values of 0.97 for the hypothalamus and 0.83 for the forebrain, indicating strong dependence between MeCP2 enrichment and mCAC. In the case of frontal cortex, an R^2 value of 0.05 revealed only modest enrichment for highly methylated regions of the genome. Using peak enrichment to indicate the relative coverage of ChIP-Seq datasets, we could identify 271,334 and 236,330 MeCP2-enriched regions in hypothalamus (Chen et al, 2015) and forebrain (Boxer et al, 2020) ChIP datasets, respectively, whereas the frontal cortex data (Gabel et al, 2015) detected only 37,817 enriched regions. This ~7-fold decrease suggests lower resolution of the frontal cortex ChIP data which may contribute to its different profile (Fig 2B) and modest correlation between MeCP2 binding and DNA methylation (Fig 2E). Regardless of differences, all three datasets failed to reveal evidence of enhanced binding of MeCP2 at $[CA]_n$, as the intensity of ChIP signal at $[CA]_n$ was approximately equivalent to that in other parts of the genome.

Based on electrophoretic mobility shift assays, it was further proposed that cooperative binding across $[CA]_n$ arrays is facilitated by the affinity of MeCP2 for non-methylated $[CA]_n$ (Ibrahim et al, 2021). This recalls early reports (Weitzel et al, 1997) that certain truncated variants of MeCP2 can bind to CA/TG-rich probes in vitro, although recent evidence failed to validate this mode of binding with full-length MeCP2 either in vitro or in vivo (Connelly et al, 2020). Ibrahim et al (2021) reported MeCP2 binding to longer non-methylated CA repeat tracts specifically $[CA]_7$. Using a pulldown assay for native MeCP2 binding in mouse brain extracts, however, we failed to detect enhanced MeCP2 binding to $[CA]_7$ compared with control DNA probes that lacked CA repeats or in which CA was part of a CAGA repeat array (Fig 3A and B). Introduction of one mC residue into the $[CA]_7$ tract significantly increased MeCP2 binding. Moreover, binding displayed a trend towards further enhancement when three more cytosines in the array were methylated. This result is compatible with a linear rather than a cooperative relationship between the amount of mCAC and MeCP2 binding. We noted that 10 mCACs in a non-repetitive probe did not appear to further enhance binding compared with four mCACs. Potential explanations for this apparent plateau include steric interference of closely proximal mCAC sites, probe length, etc. Unfortunately, the variability between experiments prevented us from exploring these alternatives quantitatively. Overall, our results fail to confirm an intrinsic affinity of MeCP2 for non-methylated $[CA]_7$ in vitro, and they suggest that addition of one mCAC motif is not sufficient to cause cooperative MeCP2 binding across a $[CA]_7$ array, as further methylation of the $[CA]_7$ probe further enhances binding.

Minor effect of CA repeats on MeCP2-mediated gene regulation

We next tested the relationship between gene expression changes in the MeCP2-deficient brain and the frequency of $[CA]_n$ repeat clusters within gene bodies, drawing on data from independent studies of mouse hypothalamus, forebrain, and cortex (Chen et al, 2015; Gabel et al, 2015; Boxer et al, 2020). To investigate the effect of enriched $[CA]_n$ repeat clusters on transcription, we asked whether significantly up- or down-regulated gene bodies were enriched for $[CA]_n$. Box plots of these gene categories failed to show obvious relationship between % $[CA]_n$ in up- or down-regulated genes compared with random non-regulated gene bodies (Fig S1A). We also plotted the percentage of all transcription unit nucleotides that belong to $[CA]_n$ arrays against the fold change in gene expression when MeCP2 is absent. This differs from a previous analysis (Ibrahim et al, 2021) by taking into account the direction of transcriptional change and testing multiple brain datasets. Again, the results showed no obvious correlation between the differing levels of $[CA]_n$ in the gene body and changes in gene expression in these brain regions (Fig 4, left panels). In contrast, we found that the number of mCAC motifs per gene, either including or excluding $[CA]_n$ repeat clusters, correlates positively with the average magnitude of gene up-regulation in the mutant brain (Fig 4, middle panels), supporting the notion that MeCP2 binding to this methylated motif restrains gene expression and confirming previous findings (Kinde et al, 2016; Lagger et al, 2017). Although gene length correlates with gene misregulation in *Mecp2 KO* (Gabel et al, 2015), a positive correlation with mCAC persisted when mCAC motifs per gene were normalized to gene length (Fig S1B). This suggests that gene length does not sufficiently explain the positive correlation with mCAC in the absence of MeCP2. The strong relationship influenced by mCAC motifs, which is unaffected by inclusion or exclusion of $[CA]_n$ arrays, is not expected if CA repeats were the primary drivers of MeCP2-mediated gene

Life Science Alliance

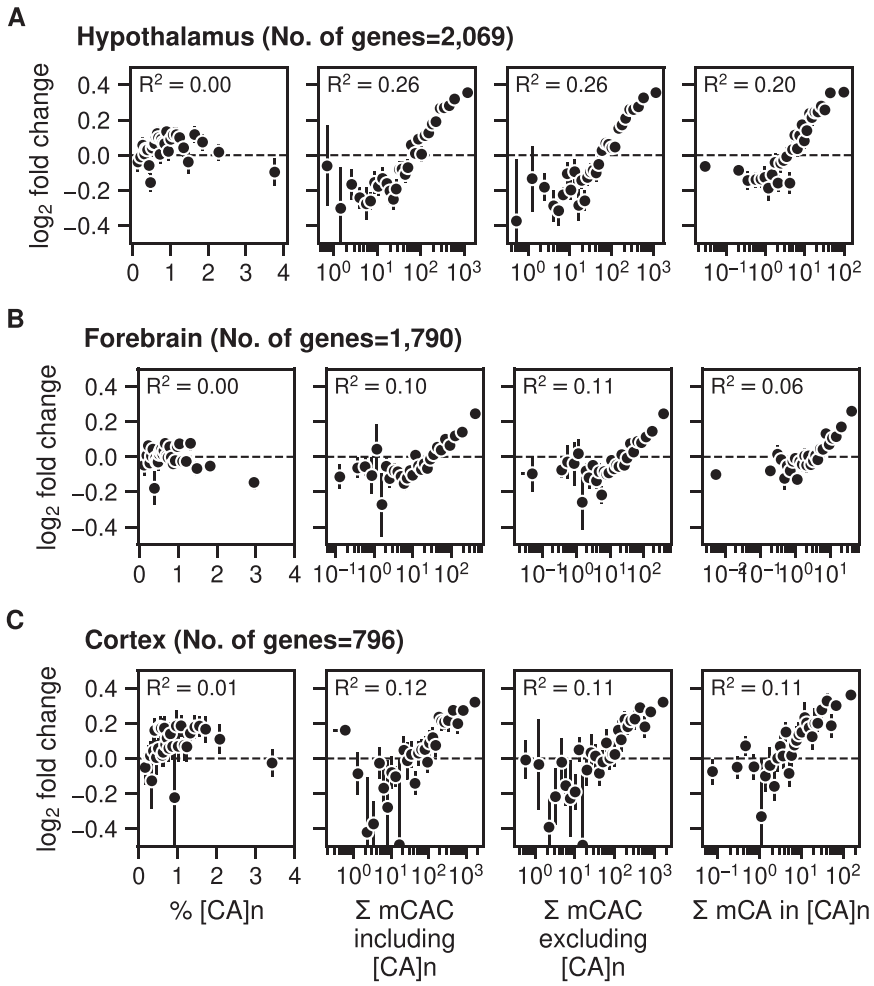


Figure 4. Gene expression increases in *Mecp2 KO* brain regions according to levels of mCAC but does not correlate with the presence of CA repeat arrays.
(A) Mean $\log_2$ fold-change in gene expression in *Mecp2 KO* versus MeCP2 *WT* in mouse hypothalamus (differentially regulated genes were defined by Padj < 0.05) (Chen et al, 2015) plotted against the percentage of CA repeats within the gene body (left panels); bins of increasing mCAC number per gene including or excluding CA repeats (Lagger et al, 2017) (centre panels); and bins of increasing mCA per gene exclusively within CA repeats (Lagger et al, 2017) (right panels). Spearman rank correlation (R^2) is calculated using unbinned values for every panel. **(B)** As described for panel (A) above but using data for mouse forebrain (differentially regulated genes were defined by Padj < 0.05) (Boxer et al, 2020). **(C)** As described for panel (A) above but using data for mouse frontal cortex (differentially regulated genes were defined by *P*-value < 0.05) (Lister et al, 2013; Gabel et al, 2015).

regulation. Because CA repeats are subject to DNA methylation at the same level as dispersed CAC motifs (see Fig 1), we expected that the presence mCAC within $[CA]_n$ tracts would correlate with gene expression. This was confirmed when the number of mCA motifs in $[CA]_n$ repeat blocks was plotted against the fold change in transcription between *Mecp2 KO* and *WT* brain regions (Fig 4, right panels). To quantify these findings, we calculated the rank correlation with $\log_2$ fold-change in gene expression for unbinned values. R^2 values were consistently higher in plots of total mCAC number per gene (including or excluding $[CA]_n$) than for % $[CA]_n$ (Fig 4, middle panels). When the level of CA methylation in $[CA]_n$ was taken into account (Fig 4, right panels), the correlation with transcriptional change did not exceed that of mCAC elsewhere in the genome. Our findings do not support the hypothesis that cytosine modification in $[CA]_n$ has a heightened impact on gene regulation.

Discussion

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

Several further considerations lead us to question the proposed link between hmC and MeCP2 (Ibrahim et al, 2021). A major reservation concerns the use of antibody detection of hmC, which provided the initial stimulus for the hypothesis of Ibrahim and colleagues (Ibrahim et al, 2021). In their experiments,

Life Science Alliance

immunoprecipitation with an antibody directed against hmC revealed prominent apparent peaks of hmC coincident with CA repeat clusters in MEFs. This result is unexpected, as levels of hmC are usually very low in dividing cultured cells compared with neurons. More importantly, others have reported "serious flaws" in the MeDIP method that lead to erroneous reporting of hmC even when both mC and hmC are known to be absent (Lentini et al, 2018). Strikingly, the resulting false positives, which account for 50–99% of regions identified as "enriched" for DNA modifications, are predominantly found at unmodified short repeat arrays, in particular $[CA]_n$. In view of this potentially serious caveat, the evidence for hmC at CA repeats in this MEF cell line must be considered provisional, pending independent biochemical validation.

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

Finally, the extreme rarity of hmCA in the brain is difficult to reconcile with its hypothetical pivotal role. A limitation of many brain methylome datasets is that only bisulfite sequence analysis was performed and therefore mC and hmC are not distinguished. It is clear, however, that although the abundance of neuronal mCA is similar to that of mCG in the brain, the vast majority of neuronal hmC is confined to hmCG (Lister et al, 2013). For example, using a non-destructive method for hmC detection, it was shown that 97.5% of hmC in excitatory neurons is in hmCG, with less than 2.5% in hmCA (Schutsky et al, 2018). This is presumably attributable to the strong preference of Tet enzymes for mCG over mCA as a substrate for mC oxidation (Hu et al, 2015; DeNizio et al, 2021). Although it is possible that hmCA plays roles in gene regulation as suggested in the cerebellum (Mellén et al, 2017), it is challenging to deconvolve its roles because of our inability to exclusively eliminate this modification from the genome. We nevertheless consider that the moderate affinity of MeCP2 for this ultra-rare motif offers an unlikely basis for comprehensive new models of MeCP2 function.

Materials and Methods

Bioinformatic analyses

Sequencing datasets

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

Table 1. Published datasets used for the analyses.

Brain tissue	Dataset	GEO accession
Hypothalamus	ChIP-Seq	GSE66868 (Chen et al, 2015)
Hypothalamus	WGBS-Seq	GSE84533 (Lagger et al, 2017)
Hypothalamus	RNA-Seq	GSE66870 (Chen et al, 2015)
Forebrain	ChIP-Seq	GSE139509 (Boxer et al, 2020)
Forebrain	WGBS-Seq	GSE128172 (Boxer et al, 2020)
Forebrain	RNA-Seq	GSE128178 (Boxer et al, 2020)
Frontal cortex	ChIP-Seq	GSE67293 (Gabel et al, 2015)
Frontal cortex	WGBS-Seq	GSE47966 (Lister et al, 2013)
Visual cortex	RNA-Seq	GSE67294 (Gabel et al, 2015)

RNA-Seq analyses

Gene expression analyses were performed for mouse brain RNA-Seq datasets (Table 1). Raw data were downloaded, mRNA expression was quantified using kallisto (Bray et al, 2016), and differential expression analysis was performed using DESeq2 (Love et al, 2014). Differentially regulated genes (significance threshold after Benjamini–Hochberg correction *P*-adjusted value < 0.05 or *P*-value < 0.05) in *Mecp2 KO* and *WT* mouse brain tissue were sorted according to the total amount of mCAC per gene body including $[CA]_n$ or excluding $[CA]_n$; total amount of mCA in $[CA]_n$ of gene body, binned into 30 equal-sized bins, and mean $\log_2$ fold-change of each bin is plotted. Error bars represent the standard error of the mean for that bin. Alternatively, the mean $\log_2$ fold-change was plotted for differentially regulated genes sorted according to the % CA repeats within the gene body and total methylation within the CA repeats.

ChIP-Seq analyses

Raw fastq reads were downloaded from GEO (Table 1) and subsequent ChIP-Seq analysis was performed on mouse genome (mm9) using snakePipes (Bhardwaj et al, 2019) v2.5.3. $\log_2$ ChIP-Seq signal over input signal and $\log_2$ wild-type ChIP-Seq signal over *Mecp2 KO* ChIP-Seq signal where available are quantified using bigWigAverageOverBed across genomic locations containing $[CA]_n$, Repeats excluding $[CA]_n$ and rest of the genome.

WGBS-Seq analyses

Processed WGBS-Seq datasets described in Table 1 were downloaded and DNA methylation ratios of individual cytosine nucleotides within CAC, CA, and CG contexts were determined across both the sense and non-sense strands. The whole genome is divided into 1 kb windows using bedtools, and DNA methylation is calculated in the 1 kb window labelled as mCAC/kb in Fig 2B. For repetitive regions, DNA methylation is calculated within the extended $[CA]_n$ or Repeats excluding $[CA]_n$ region to match the length of repeat regions equivalent to 1 kb. Mean DNA methylation levels for different

Life Science Alliance

Table 2. Oligonucleotide sequences for the probes used in the pulldown assay.

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

B, biotin, m, methyl group. All molecules were annealed to the appropriate methylated or non-methylated reverse oligonucleotide.

DNA sequence contexts are calculated for cytosines with adequate sequence coverage. For $[CA]_n$ and $[CAN]_n$, all cytosines across both strands within the genomic loci of respective repeats are considered. Because of coverage differences between the bisulfite datasets, we set the cytosine coverage thresholds for the hypothalamus (Lagger et al, 2017) and forebrain (Boxer et al, 2020) at five compared with more highly covered cortex (Lister et al, 2013) at at least 10 reads for every cytosine. These thresholds enable reliable estimates of average DNA methylation.

$[CA]_n$, $[CAN]_n$, and Repeats excluding $[CA]_n$

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

CAC occurrences

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

Reproducibility

Source code to reproduce all the analysis and figures is available on the GitHub repository (https://github.com/kashyapchhatbar/MeCP2_2022_manuscript) and archived at Zenodo (DOI: 10.5281/zenodo.6997675).

Pulldown assay for MeCP2 binding to DNA

This assay was performed as described previously (Piccolo et al, 2019) with the following modifications. Biotin end-labelled double-strand synthetic oligonucleotides (2 μg) described in Table 2 were coupled to M280-streptavidin Dynabeads according to manufacturer's instructions (Invitrogen). Bead-DNA complex was then co-incubated at 4°C for 1.5 h with nuclear protein (10 μg). Nuclear extracts from mouse brain (0.42 M salt) were prepared as described (Mellén et al, 2017) and dialysed back into a solution containing 0.15 M NaCl. After extensive washing, bead-bound proteins were eluted in Laemmli buffer (Sigma-Aldrich) and resolved on a 4–20% SDS-polyacrylamide gel (NEB). The presence of MeCP2 was assayed by Western blotting using anti-MeCP2 monoclonal antibody M6818 (Sigma-Aldrich) using IR dye as a secondary antibody (IRDye 800CW donkey anti-mouse; LI-COR Biosciences). Triplicate assays were scanned then quantified using a LI-COR Odyssey CLx machine and software.

Supplementary Information

Supplementary Information is available at https://doi.org/10.26508/lsa.202201522.

Acknowledgements

We thank Raphaël Pantier and Matthew Lyst for providing critical feedback on the work. This work was supported by Wellcome Centre grant 107930/Z/15/Z to A Bird, who is also a member of the Simons Initiative for the Developing Brain. K Chhatbar was supported by a staff scholarship from the College of Science and Engineering, University of Edinburgh.

Author Contributions

K Chhatbar: conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, and writing—original draft, review, and editing.
J Connelly: conceptualization, validation, investigation, visualization, methodology, and writing—original draft, review, and editing.
S Webb: data curation, software, formal analysis, and writing—review and editing.
S Kriaucionis: conceptualization, data curation, supervision, investigation, methodology, and writing—original draft, review, and editing.
A Bird: conceptualization, resources, supervision, funding acquisition, validation, investigation, visualization, project administration, and writing—original draft, review, and editing.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Ayata P (2013) Decoding 5HMC as an active chromatin mark in the brain and its link to Rett syndrome. *Student Theses and Dissertations* 238. https://digitalcommons.rockefeller.edu/student_theses_and_dissertations/238

Bhardwaj V, Heyne S, Sikora K, Rabbani L, Rauer M, Kilpert F, Richter AS, Ryan DP, Manke T (2019) snakePipes: Facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics* 35: 4757–4759. doi:10.1093/bioinformatics/btz436

Boxer LD, Renthal W, Greben AW, Whitwam T, Silberfeld A, Stroud H, Li E, Yang MG, Kinde B, Griffith EC, et al (2020) MeCP2 represses the rate of transcriptional initiation of highly methylated long genes. *Mol Cell* 77: 294–309.e9. doi:10.1016/j.molcel.2019.10.032

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34: 525–527. doi:10.1038/nbt.3519

Chen L, Chen K, Lavery LA, Baker SA, Shaw CA, Li W, Zoghbi HY (2015) MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proc Natl Acad Sci U S A* 112: 5509–5514. doi:10.1073/pnas.1505909112

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. *Proc Natl Acad Sci U S A* 116: 14995–15000. doi:10.1073/pnas.1903549116

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

de Mendoza A, Poppe D, Buckberry S, Pflueger J, Albertin CB, Daish T, Bertrand S, de la Calle-Mustienes E, Gómez-Skarmeta JL, Nery JR, et al (2021) The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* 5: 369–378. doi:10.1038/s41559-020-01371-2

DeNizio JE, Dow BJ, Serrano JC, Ghanty U, Drohat AC, Kohli RM (2021) TET-TDG active DNA demethylation at CpG and non-CpG sites. *J Mol Biol* 433: 166877. doi:10.1016/j.jmb.2021.166877

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, Greenberg ME (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* 522: 89–93. doi:10.1038/nature14319

Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, 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. doi:10.1038/nn.3607

Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, Zhang X, Cheng X (2012) Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* 40: 4841–4849. doi:10.1093/nar/gks155

Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, Limouse C, Halabian R, Wojenski L, O'Neill RJ (2022) From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* 376.6588: eabk3112.

Hu L, Lu J, Cheng J, Rao Q, Li Z, Hou H, Lou Z, Zhang L, Li W, Gong W, et al (2015) Structural insight into substrate preference for TET-mediated oxidation. *Nature* 527: 118–122. doi:10.1038/nature15713

Ibrahim A, Papin C, Mohideen-Abdul K, Le Gras S, Stoll I, Bronner C, Dimitrov S, Klaholz BP, Hamiche A (2021) MeCP2 is a microsatellite binding protein that protects CA repeats from nucleosome invasion. *Science* 372: eabd5581. doi:10.1126/science.abd5581

Kinde B, Wu DY, Greenberg ME, Gabel HW (2016) DNA methylation in the gene body influences MeCP2-mediated gene repression. *Proc Natl Acad Sci U S A* 113: 15114–15119. doi:10.1073/pnas.1618737114

Lavery LA, Ure K, Wan Y-W, Luo C, Trostle AJ, Wang W, Jin H, Lopez J, Lucero J, Durham MA, et al (2020) Losing Dnmt3a dependent methylation in inhibitory neurons impairs neural function by a mechanism impacting Rett syndrome. *Elife* 9: e52981. doi:10.7554/elife.52981

Lagger S, Connelly JC, Schweikert G, Webb S, Selfridge J, Ramsahoye BH, Yu M, He C, Sanguinetti G, Sowers LC, et al (2017) MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet* 13: e1006793. doi:10.1371/journal.pgen.1006793

Lei M, Tempel W, Chen S, Liu K, Min J (2019) Plasticity at the DNA recognition site of the MeCP2 mCG-binding domain. *Biochim Biophys Acta Gene Regul Mech* 1862: 194409. doi:10.1016/j.bbagrm.2019.194409

Lentini A, Lagerwall C, Vikingsson S, Mjoseng HK, Douvlataniotis K, Vogt H, Green H, Meehan RR, Benson M, Nestor CE (2018) A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nat Methods* 15: 499–504. doi:10.1038/s41592-018-0038-7

Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, Bird A (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. *Cell* 69: 905–914. doi:10.1016/0092-8674(92)90610-o

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, et al (2013) Global epigenomic reconfiguration during mammalian brain development. *Science* 341: 1237905. doi:10.1126/science.1237905

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550. doi:10.1186/s13059-014-0550-8

Mellén M, Ayata P, Heintz N (2017) 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc Natl Acad Sci U S A* 114: E7812–E7821. doi:10.1073/pnas.1708044114

Piccolo FM, Liu Z, Dong P, Hsu C-L, Stoyanova EI, Rao A, Tjian R, Heintz N (2019) MeCP2 nuclear dynamics in live neurons results from low and high affinity chromatin interactions. *Elife* 8: e51449. doi:10.7554/elife.51449

Life Science Alliance

Schutsky EK, DeNizio JE, Hu P, Liu MY, Nabel CS, Fabyanic EB, Hwang Y, Bushman FD, Wu H, Kohli RM (2018) Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat Biotechnol* 36: 1083–1090. doi:10.1038/nbt.4204

Tillotson R, Cholewa-Waclaw J, Chhatbar K, Connelly JC, Kirschner SA, Webb S, Koerner MV, Selfridge J, Kelly DA, De Sousa D, et al (2021) Neuronal non-CG methylation is an essential target for MeCP2 function. *Mol Cell* 81: 1260–1275.e12. doi:10.1016/j.molcel.2021.01.011

Valinluck V, Tsai H-H, Rogstad DK, Burdzy A, Bird A, Sowers LC (2004) Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* 32: 4100–4108. doi:10.1093/nar/gkh739

Weitzel JM, Buhrmester H, Strätling WH (1997) Chicken MAR-binding protein ARBP is homologous to rat methyl-CpG-binding protein MeCP2. *Mol Cell Biol* 17: 5656–5666. doi:10.1128/mcb.17.9.5656

License: This article is available under a Creative Commons License (Attribution 4.0 International, as described at https://creativecommons.org/licenses/by/4.0/).

4.2.3 Conclusion

Our findings indicate that CA repeats undergo CAC methylation at levels consistent with the surrounding genome, and MeCP2-mediated gene expression modulation is expected to occur even in the absence of CA repeats in the mouse genome. A comparative analysis of mouse and human genomes reveals that the mouse genome is more enriched in CA repeat sequences than the human genome. Given that satellite DNA is a major component of chromocenters, this finding aligns with recent observations (Pantier et al. 2024) that MeCP2 binding is diffused in most mammals, including humans where lack of functional MeCP2 causes Rett Syndrome, whereas it is concentrated at chromocenters in mouse cells.

Chapter 5

Influence of DNA base composition and SALL4 on gene regulation

As outlined in Section 3.2, sequence composition is a fundamental but often overlooked feature of the genome. We test the hypothesis that SALL4, a key transcriptional regulator, reads and responds to AT-rich sequences to influence transcription. By integrating computational modelling with functional genomics data, we examine whether gene body base composition correlates with SALL4-mediated repression. The experiments performed to address this hypothesis, along with their results, are presented as they appear in my co-authored peer-reviewed manuscript (Pantier, Chhatbar, et al. 2021) [1].

5.1 Introduction

Mammalian genomes exhibit non-random distributions of A/T and G/C base pairs, forming large compositional domains that coincide with various functional elements. While the significance of GC-rich regions and CGIs in transcriptional regulation is well established, the role of AT-rich sequences remains less understood. We test the hypothesis that base composition can be read as a signal, influencing gene expression through its interaction with chromatin-associated factors.

This study explores how base composition contributes to transcriptional regulation, focusing on its impact on gene repression. Using a combination of linear regression modelling and functional genomics data, we quantify the relationship between gene body base composition and transcriptional repression, demonstrating how sequence composition influences gene regulation.

5.2 SALL4 controls cell fate in response to base composition

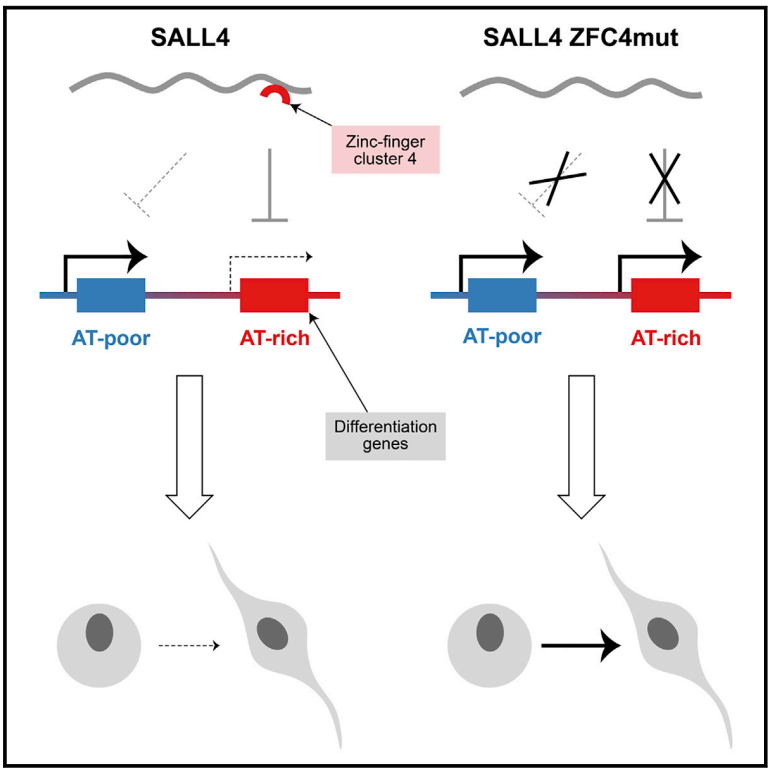
[1] This manuscript will appear as is where all computational biology analysis is performed by me, the laboratory experiments have been performed by Raphaël Pantier and Timo Quante, and the material is collaboratively written by myself, Raphaël Pantier and Adrian Bird.

Molecular Cell

Article

SALL4 controls cell fate in response to DNA base composition

Graphical Abstract



Authors

Raphaël Pantier, Kashyap Chhatbar, Timo Quante, ..., Michiel Vermeulen, Jim Selfridge, Adrian Bird

Correspondence

a.bird@ed.ac.uk

In Brief

The mammalian genome comprises domains with varying ratios of A/T to G/C base pairs. The significance of mosaic base composition was unknown, but we discovered that the stem cell protein SALL4 reads A/T content to inhibit the expression of many genes, thereby stabilizing the pluripotent state.

Highlights

- SALL4 binds to short AT-rich motifs via its C-terminal zinc-finger cluster, ZFC4
- SALL4 ZFC4 represses early differentiation genes in proportion to their AT content
- Mutating ZFC4 leads to abnormal neuronal differentiation and embryonic lethality
- The other zinc-finger clusters ZFC1 and ZFC2 seem dispensable in embryonic stem cells

Pantier et al., 2021, Molecular Cell *81*, 845–858
February 18, 2021 © 2020 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.molcel.2020.11.046

CellPress

Molecular Cell

CellPress
OPEN ACCESS

Article

SALL4 controls cell fate in response to DNA base composition

Raphaël Pantier,[1,4] Kashyap Chhatbar,[1,3,4] Timo Quante,[1,4] Konstantina Skourti-Stathaki,[1] Justyna Cholewa-Waclaw,[1] Grace Alston,[1] Beatrice Alexander-Howden,[1] Heng Yang Lee,[1] Atlanta G. Cook,[1] Cornelia G. Spruijt,[2] Michiel Vermeulen,[2] Jim Selfridge,[1] and Adrian Bird[1,5,*]

[1]The Wellcome Centre for Cell Biology, University of Edinburgh, Michael Swann Building, Max Born Crescent, The King's Buildings, Edinburgh EH9 3BF, UK

[2]Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Oncode Institute, Radboud University Nijmegen, Nijmegen, the Netherlands

[3]Informatics Forum, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK

[4]These authors contributed equally

[5]Lead contact

*Correspondence: a.bird@ed.ac.uk

https://doi.org/10.1016/j.molcel.2020.11.046

SUMMARY

Mammalian genomes contain long domains with distinct average compositions of A/T versus G/C base pairs. In a screen for proteins that might interpret base composition by binding to AT-rich motifs, we identified the stem cell factor SALL4, which contains multiple zinc fingers. Mutation of the domain responsible for AT binding drastically reduced SALL4 genome occupancy and prematurely upregulated genes in proportion to their AT content. Inactivation of this single AT-binding zinc-finger cluster mimicked defects seen in *Sall4* null cells, including precocious differentiation of embryonic stem cells (ESCs) and embryonic lethality in mice. In contrast, deletion of two other zinc-finger clusters was phenotypically neutral. Our data indicate that loss of pluripotency is triggered by downregulation of SALL4, leading to de-repression of a set of AT-rich genes that promotes neuronal differentiation. We conclude that base composition is not merely a passive byproduct of genome evolution and constitutes a signal that aids control of cell fate.

INTRODUCTION

A/T and G/C base pairs are nonrandomly distributed within mammalian genomes, forming large and relatively homogeneous AT-rich or GC-rich regions that usually encompass several genes together with their intergenic sequences. Base compositional domains are often evolutionarily conserved (Bernardi et al., 1985; Holmquist, 1989; Bickmore and Sumner, 1989; Costantini et al., 2009) and coincide with other genomic features (Bickmore and van Steensel, 2013), including early/late-replicating regions (Hiratani et al., 2009; Holmquist et al., 1982), lamina-associated domains (Meuleman et al., 2013), and topologically associating domains (Jabbari and Bernardi, 2017). Despite these interesting correlations, it is unclear whether conserved AT-rich and GC-rich domains are passive byproducts of evolution or DNA base composition can play an active biological role (Eyre-Walker and Hurst, 2001; Duret et al., 2006; Arhondakis et al., 2011). Exemplifying this second hypothesis, CpG islands represent conserved GC-rich domains (Bird, 1986) that are specifically bound by proteins recognizing unmethylated "CG" dinucleotides (Lee et al., 2001; Voo et al., 2000) to modulate chromatin structure and regulate gene expression (Thomson et al., 2010; Blackledge et al., 2010; Farcas et al., 2012; Wu et al., 2013; He et al., 2013).

Here, we tested the hypothesis that AT-rich DNA can be interpreted by specific proteins that recognize short AT-rich motifs whose frequency tracks fluctuations in base composition across the genome (Quante and Bird, 2016). To identify novel AT-binding proteins, we utilized a DNA pull-down mass spectrometry screen in mouse embryonic stem cells (ESCs), which are pluripotent and can be differentiated in culture. As a top hit, we identified SALL4, which is a multi-zinc-finger protein that restrains differentiation of ESCs (Yuri et al., 2009; Miller et al., 2016) and participates in several physiological processes, including neuronal development (Böhm et al., 2008; Sakaki-Yumoto et al., 2006; Tahara et al., 2019), limb formation (Akiyama et al., 2015; Koshiba-Takeuchi et al., 2006), and gametogenesis (Chan et al., 2017; Hobbs et al., 2012; Xu et al., 2017; Yamaguchi et al., 2015). Deletion of the *Sall4* gene leads to embryonic lethality shortly after implantation (Sakaki-Yumoto et al., 2006; Elling et al., 2006; Warren et al., 2007). In humans, failure of SALL4 function is the cause of two severe developmental disorders: the recessive genetic disorder Okihiro syndrome (Al-Baradie et al., 2002; Kohlhase et al., 2002) and embryopathies due to

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

CellPress OPEN ACCESS

Molecular Cell
Article

treatment during pregnancy with the drug thalidomide (Donovan et al., 2018; Matyskiela et al., 2018). Despite its biological and biomedical importance, the molecular functions of SALL4 are incompletely understood. The extreme N terminus interacts specifically with the nucleosome remodeling and deacetylase (NuRD) co-repressor complex and can account for the transcriptional repression caused by SALL4 recruitment to reporter genes (Lauberth and Rauchman, 2006; Lu et al., 2009). In addition, there is evidence that the zinc-finger clusters bind to DNA (Sakaki-Yumoto et al., 2006; Xiong et al., 2016) or protein partners (Hobbs et al., 2012; Tanimura et al., 2013), though their precise developmental roles are unclear. The present work demonstrates that many of the defects seen in *Sall4* null ESCs, including precocious differentiation, are mimicked by inactivation of a single zinc-finger cluster that interacts with AT-rich motifs. We go on to show that the ability of SALL4 to sense DNA base composition is essential to restrain transcription of genes that promote differentiation.

RESULTS

A screen for AT-binding proteins in ESCs identifies SALL4

To identify proteins able to sense base composition, we used a DNA affinity purification approach coupled with stable isotope labeling with amino acids in cell culture (SILAC) quantitative mass spectrometry (Spruijt et al., 2013a, 2013b). Mouse ESC protein extracts were mixed with double-stranded DNA probes carrying variable runs of 5 base pairs composed only of A or T nucleotides (AT-1 and AT-2). As a negative control, matched probes with AT runs interrupted by G or C nucleotides were used as bait (Ctrl-1 and Ctrl-2). To robustly assess DNA-binding specificity, experiments were performed both in the "forward" (heavy-labeled AT probe versus light-labeled Ctrl-probe) and "reverse" (heavy-labeled Ctrl-probe versus light-labeled AT probe) orientations, which were considered as biological replicates. Proteins that bind specifically to AT runs show a high ratio in the forward experiments (Figure 1A, x axes) and a low ratio in the reverse experiments (Figure 1A, y axes). Mass spectrometry identified a consistent set of AT-binding proteins that largely overlapped between replicate experiments (Figure S1A) and between unrelated AT-rich probes (Figure S1B). High-confidence hits included proteins with well-characterized AT-binding domains such as AT hooks (Aravind and Landsman, 1998; Filarsky et al., 2015) (HMGA1, HMGA2, PRR12, and BAZ2A) and "AT-rich interaction domains" (Patsialou et al., 2005) (ARID3A, ARID3B, and ARID5B), thereby validating the screen (Figure 1A; Table S1). Three Spalt-like (SALL) family proteins (Sweetman and Münsterberg, 2006) (SALL1, SALL3, and SALL4) and most components of the NuRD complex (Tong et al., 1998; Wade et al., 1998; Xue et al., 1998; Zhang et al., 1998) were also recovered (Figure 1A). The most consistently enriched protein in our mass spectrometry screen was SALL4, whose AT binding we confirmed by western blot analysis using a variety of probes with one (AT-3) or more AT runs (Figure S1C). Considering their reported interaction with NuRD (Lauberth and Rauchman, 2006), we suspected that SALL proteins might be responsible for recruitment of this co-repressor complex to AT-rich DNA. To test this, we used extracts from mouse ESCs in which the *Sall4* gene is disrupted (*Sall4* knockout [*S4KO*] ESCs) (Sakaki-Yumoto et al., 2006; Miller et al., 2016). As predicted, recovery of NuRD components by AT-rich DNA was greatly reduced compared to *wild type* (*WT*) in the absence of SALL4 (Figure 1B).

SALL4 binds to short AT-rich motifs via C2H2 zinc-finger cluster 4 (ZFC4)

Mammalian genomes encode four SALL family proteins (SALL1–SALL4), which each contain several clusters of C2H2 zinc fingers. Based on similarities in amino acid sequence between family members, the clusters are classified as ZFC1–ZFC4 (Figure 1C). SALL1, SALL3, and SALL4 all possess ZFC4 (Figure S1D), but SALL2 lacks this domain and was not recovered in our screen for AT-binding proteins. ZFC4 of both SALL1 and SALL4 was previously shown to interact with AT-rich heterochromatin in transfection assays (Sakaki-Yumoto et al., 2006; Yamashita et al., 2007), suggesting that it might be responsible for AT binding. To further characterize this domain, we used CRISPR-Cas9 to either delete ZFC4 (*ZFC4Δ*) or mutate two residues (T919D and N922A; mutated residues are shown in red) that we predicted would be involved in DNA binding (*ZFC4mut*) (Figure 1D). Homozygous mouse ESC lines expressing both mutated SALL4 proteins were obtained (Figure S1E), both of which retained the ability to interact with NuRD components by co-immunoprecipitation (Figure S1F). The interaction of SALL4 ZFC4mut or ZFC4Δ proteins with AT-rich sequences was drastically reduced (>10-fold) by inactivation of ZFC4, as assessed by the DNA pull-down assay (Figure 1E). This strongly suggests that the ZFC4 domain of SALL4 is primarily responsible for pull-down by AT-rich DNA. We next explored the *in vivo* DNA-binding properties of SALL4 ZFC4 in our mutant ESC lines. Heterochromatic foci, identified by DAPI staining in mouse cells, contain a high concentration of AT-rich satellite DNA (Matsuda and Chapman, 1991; Cerda et al., 1999; Guenatri et al., 2004) and therefore provide a low-resolution cellular assay for preferential AT binding. Immunostaining with a SALL4 antibody recognizing a preserved epitope in the two mutant proteins revealed a striking loss of ZFC4mut and ZFC4Δ protein localization at DAPI-dense foci (Figure 1F), further confirming that this zinc-finger cluster is necessary for AT targeting.

To define the sequence preference of the purified ZFC4 domain, we performed systematic evolution of ligands by exponential enrichment (SELEX) coupled with high-throughput sequencing (HT-SELEX), whereby a library of initially random DNA sequences immobilized on beads was subjected to repeated cycles of binding and amplification (Jolma et al., 2010; Nitta et al., 2015). After 0, 1, 3, or 6 cycles, DNA was analyzed by high-throughput sequencing to detect enriched motifs. For comparison, we performed HT-SELEX on other SALL4 zinc-finger clusters (ZFC1 and ZFC2) (Figure 2A) and also included a sample without added proteins to control for PCR bias. Strikingly, with increasing cycles of ZFC4 binding, the base composition of the whole library gradually shifted toward higher AT content, but this effect was not seen with ZFC1, ZFC2, or the negative control (Figure 2B). Progressive A/T motif enrichment was also apparent for ZFC4 alone

Molecular Cell
Article

CellPress
OPEN ACCESS

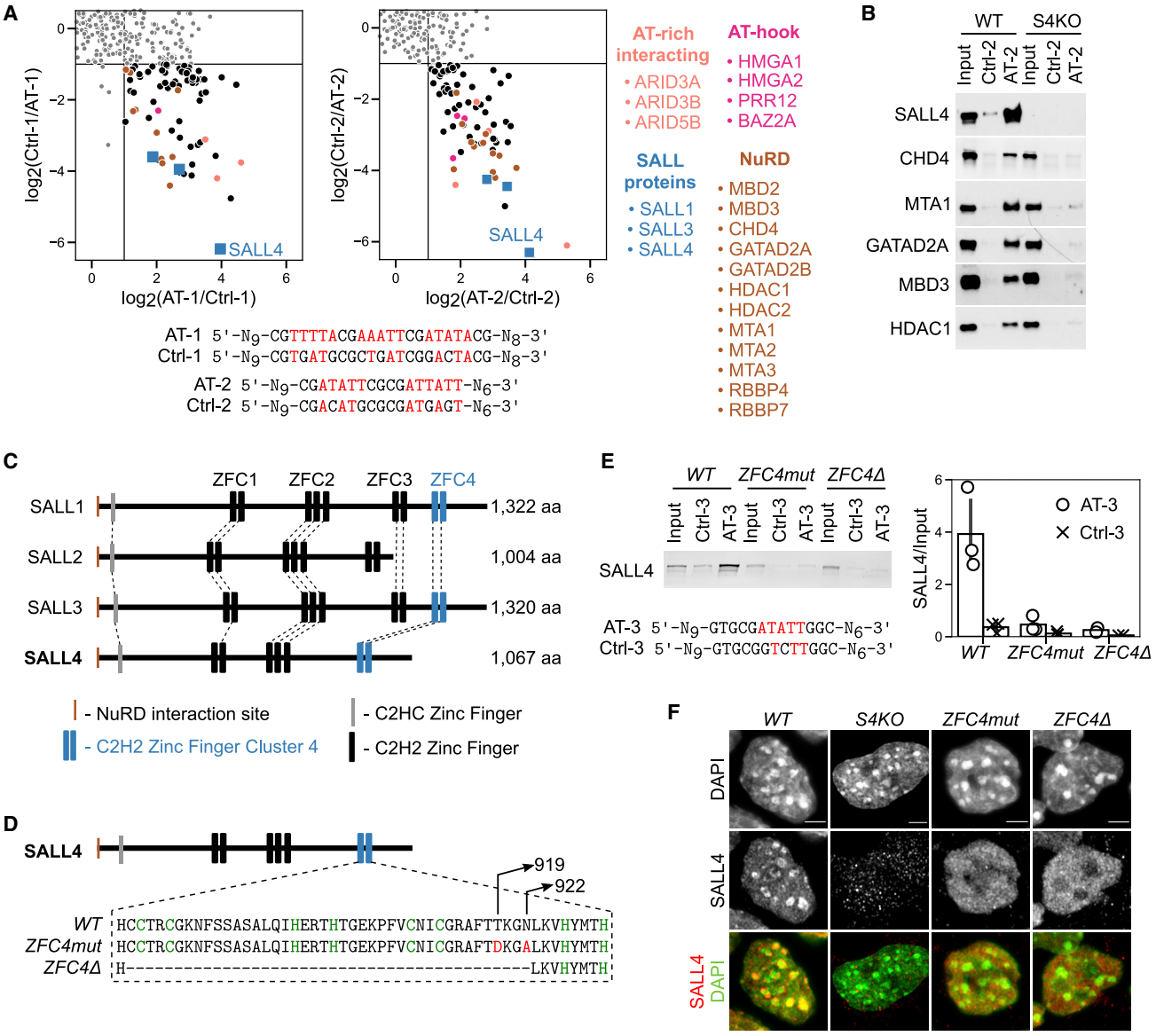


Figure 1. Identification of novel AT-binding proteins in ESCs by DNA pull-down mass spectrometry

(A) Scatterplots representing SILAC-based DNA affinity purifications, comparing AT-rich DNA probes (AT-1 and AT-2) with control probes having interrupted AT runs (Ctrl-1 and Ctrl-2). The ratio of the quantified proteins in the forward experiment (Heavy-AT/Light-Ctrl) is plotted on the x axis, and the ratio for the same proteins in the reverse experiment (Heavy-Ctrl/Light-AT) is plotted on the y axis. Proteins were considered to be specific interactors when showing at least a 2-fold ratio in both the forward and reverse experiments. These proteins cluster in the bottom right quadrant.

(B) DNA pull-down with AT-rich (AT-2) or control (Ctrl-2) probes followed by western blot analysis for SALL4 and NuRD components using *wild-type* (*WT*) or *Sall4* knockout (*S4KO*) ESC protein extracts.

(C) Protein alignment of mouse SALL family members indicating conserved protein domains, including C2H2 zinc-finger clusters (ZFC1–ZFC4).

(D) Diagram showing the mutations or deletion introduced within SALL4 ZFC4 by CRISPR-Cas9.

(E) DNA pull-down with AT-rich (AT-3) or control (Ctrl-3) probe followed by western blot analysis for SALL4 using *WT* or *Sall4 ZFC4mut/Δ* ESC protein extracts. SALL4 levels were quantified and normalized to input. Data points indicate independent replicate experiments and error bars standard deviation.

(F) SALL4 immunofluorescence in the indicated ESC lines. DNA was stained with DAPI, showing dense clusters of AT-rich pericentric chromatin. Scale bars, 3 μm.

See also Figure S1 and Table S1.

(Figure 2C). To determine the minimum number of A/T base pairs required for enrichment, we separated the oligomers (k-mers) into different groups (Figure 2C). Enrichment was only detectable in k-mers containing 4 or more A/T base pairs, suggesting that this is the minimum target sequence. After 6 cycles, the most enriched SELEX motif was "ATATT" (Figure S2A),

CellPress OPEN ACCESS

Molecular Cell
Article

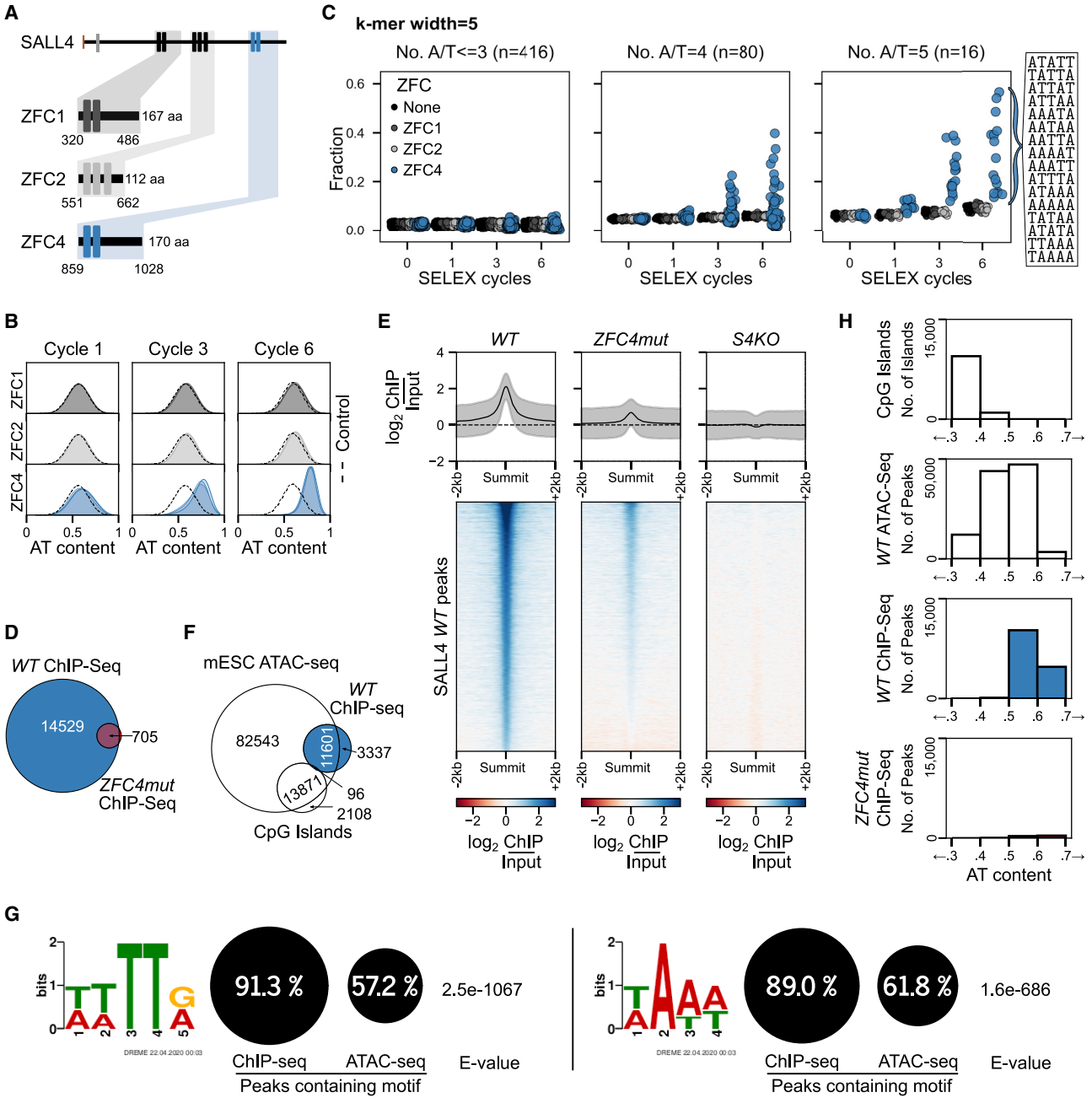


Figure 2. Characterization of SALL4 C2H2 zinc-finger cluster 4 (ZFC4) DNA binding *in vitro* and *in vivo*

(A) SALL4 ZFC1, ZFC2, and ZFC4 protein fragments used for *in vitro* HT-SELEX experiments. A sample without added protein was used as a negative control.

(B) Base composition of HT-SELEX libraries following 1, 3, and 6 cycles with ZFC1 (dark gray), ZFC2 (light gray) and ZFC4 purified proteins (blue).

(C) Relative enrichment of 5-mer motifs categorized by AT content at cycles 0, 1, 3, and 6 of HT-SELEX with SALL4 ZFC1 (dark gray), ZFC2 (light gray), ZFC4 (blue), and negative control (black) samples. All 16 possible A/T 5-mer motifs are ordered according to their enrichment at cycle 6 of ZFC4 HT-SELEX.

(D) Venn diagram showing the overlap of SALL4 ChIP-seq peaks between *WT* and *ZFC4mut* ESCs.

(E) Profile plot and heatmap showing SALL4 ChIP-seq signal at SALL4 *WT* ChIP-seq peaks in the indicated cell lines.

(F) Venn diagram showing the overlap of SALL4 ChIP-seq peaks detected in *WT* ESCs with ATAC-seq peaks (accessible chromatin) and CpG islands.

(G) Results from *de novo* motif analysis at SALL4 *WT* ChIP-seq peaks (summit ±250 bp) showing the relative frequency of each DNA motif and its associated E-value. ATAC-seq peaks were used as a control for regions of accessible chromatin.

(H) Analysis of the DNA base composition surrounding SALL4 ChIP-seq peaks (summit ±250 bp) in *WT* (blue) and *ZFC4mut* (red) ESCs. CpG islands and ATAC-seq peaks coincide with regions of accessible chromatin and are shown for comparison.

See also Figure S2.

Molecular Cell
Article

CellPress
OPEN ACCESS

which also corresponds to the preferred sequence identified by DNA pull-down using all possible combinations of AT 5-mers (Figure S2B). However, this is one of several target sequences, as multiple other AT-rich sequences had similar enrichment scores (Figures 2B and S2A). All 16 possible A/T 5-mers are enriched more than G/C containing 5-mers, with k-mers containing TATA among the least favored AT-rich motifs for ZFC4 binding. We conclude that ZFC4 targets a broad range of short motifs that are composed only of A and T, whereas ZFC1 and ZFC2 showed no apparent DNA sequence specificity.

ZFC4 mutation drastically reduces SALL4 chromatin binding *in vivo*

To assess the influence of ZFC4 on SALL4 chromatin occupancy *in vivo*, we performed chromatin immunoprecipitation sequencing (ChIP-seq) using two anti-SALL4 antibodies (one monoclonal and one polyclonal) recognizing a C-terminal epitope that is distant from C2H2 zinc-finger clusters. We first determined antibody specificity (Kidder et al., 2011; Landt et al., 2012; Uhlen et al., 2016) by assessing SALL4 ChIP signal in *S4KO* ESCs as a negative control. Over 15,000 nonspecific ChIP-seq peaks were observed with the polyclonal anti-SALL4 antibody compared with only 280 peaks for the monoclonal antibody (Figures S2C and S2D). We therefore analyzed exclusively data obtained with the anti-SALL4 monoclonal antibody, considering only ChIP-seq peaks that were consistent between independent replicate experiments in *WT* or *ZFC4mut* ESCs (Figure S2E). In agreement with its reported localization at enhancers (Miller et al., 2016; Xiong et al., 2016), we observed that SALL4 ChIP-seq peaks in *WT* cells were enriched in the histone marks H3K27ac and H3K4me1 (Chronis et al., 2017), which typically mark these genomic sites (Figure S2F). Strikingly, *ZFC4mut* cells lost ~95% of ChIP-seq peaks compared to *WT* (Figure 2D). Heatmaps confirmed the depletion of SALL4 peaks, although we noted a small amount of bound ZFC4mut at a subset of *WT* binding sites (Figure 2E).

We compared *WT* SALL4-binding sites as a whole with regions of open chromatin identified by assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), which detects accessible DNA, including enhancers and promoters. SALL4 peaks largely coincide with a subset of ATAC-seq peaks, while avoiding CpG island promoters (Figure 2F). The AT-binding specificity of SALL4 suggests that this protein might preferentially associate with open chromatin sites that are more AT rich than average. The complete absence of SALL4 at ATAC-seq peaks within CpG islands (Figure 2F), within which runs of As and Ts are rare, is compatible with this notion. To quantify this effect, we used *de novo* motif analysis to determine whether SALL4 peaks were consistent with a bias toward AT-rich motifs. First, by seeking recurrent motifs (<8 base pairs) coincident with SALL4 peaks, we identified short AT-rich motifs that were highly enriched at the majority (~90%) of SALL4-binding sites compared with lower levels of enrichment (~60%) in open chromatin generally (Figure 2G). As a second approach, we determined the base composition at SALL4-bound regions by analyzing the DNA sequences surrounding SALL4 ChIP-seq peak summits (±250 bp). SALL4 binding sites are relatively AT rich (50%–70% AT) (Figure 2H) compared with ATAC-seq peaks as a whole (40%–60% AT) (Figure 2H). Taken together, the data suggest that AT-motif binding is responsible for the presence of SALL4 at a subset of open chromatin sites.

SALL4 ZFC4 represses the expression of early differentiation genes in a base-composition-dependent manner

To determine whether SALL4 binding to AT-rich DNA causes gene expression changes that correlate with base composition, we performed RNA sequencing (RNA-seq) in *WT*, *ZFC4mut*, *ZFC4Δ*, and *S4KO* ESCs. *Sall4* gene knockout resulted in the dysregulation of several thousand genes (Figure 3A). Both ZFC4 mutations caused the dysregulation of fewer genes, many of which overlapped with those affected in *S4KO* cells (Figure 3A). To test the relationship between AT composition and gene expression, genes differentially regulated in both *ZFC4mut* and *ZFC4Δ* ESCs (Figure 3A, red filling) were divided into five equal categories according to AT content across the entire transcription unit (Figure 3B), and the level and direction of transcriptional change was compared between them. In agreement with our hypothesis, genes differentially regulated in *ZFC4mut/Δ* cells showed progressively increased upregulation with rising AT content (Figure 3C). To quantify the strength of the relationship between AT content and gene expression, we fitted a linear regression model and calculated coefficient estimates. This independent approach, which reveals the variation in gene expression that can be attributed to base composition, confirmed that the positive relationship between AT content and upregulation in the ZFC4 mutants is significant (false discovery rate [FDR] < 0.01; see STAR methods; Table S2; Figure S3A). In contrast, genes differentially regulated in *S4KO,* but not in either of the *ZFC4* mutant ESCs (Figure 3A, gray filling), showed a nonsignificant correlation (FDR > 0.01) and an effect size close to zero (Figures 3D, S3B, and S3C). The results show that the subset of SALL4-regulated genes that is dependent on ZFC4 is repressed in pluripotent cells according to the AT richness of their genomic setting.

To further test the hypothesis that AT binding by ZFC4 mediates repression according to base composition, we examined the reverse situation of SALL4 overexpression on transcription. This was performed by expressing SALL4, or as a negative control enhanced green fluorescent protein (EGFP), from a doxycycline-inducible promoter following random integration of expression constructs in *S4KO* ESCs (Figure 3E). After 48 h of induction, SALL4 was robustly overexpressed in these cells (Figures S3D and S3E). To characterize the effect of SALL4 reexpression on transcription, we performed RNA-seq on induced (+Dox) and noninduced (−Dox) cell lines (Figure 3F). As expected, gene expression changes in cells overexpressing SALL4 were anticorrelated with expression changes seen in *S4KO* cells (Figure S3F). Separation of differentially expressed genes into categories according to their AT content as before revealed that SALL4 expression caused transcriptional repression that was strikingly proportional to the base composition of the affected genes (Figures 3G and S3G). A similar relationship was observed when looking at genes differentially regulated in *ZFC4mut/Δ* ESCs (Figure 3H). Linear regression analysis again confirmed the

CellPress OPEN ACCESS

Molecular Cell
Article

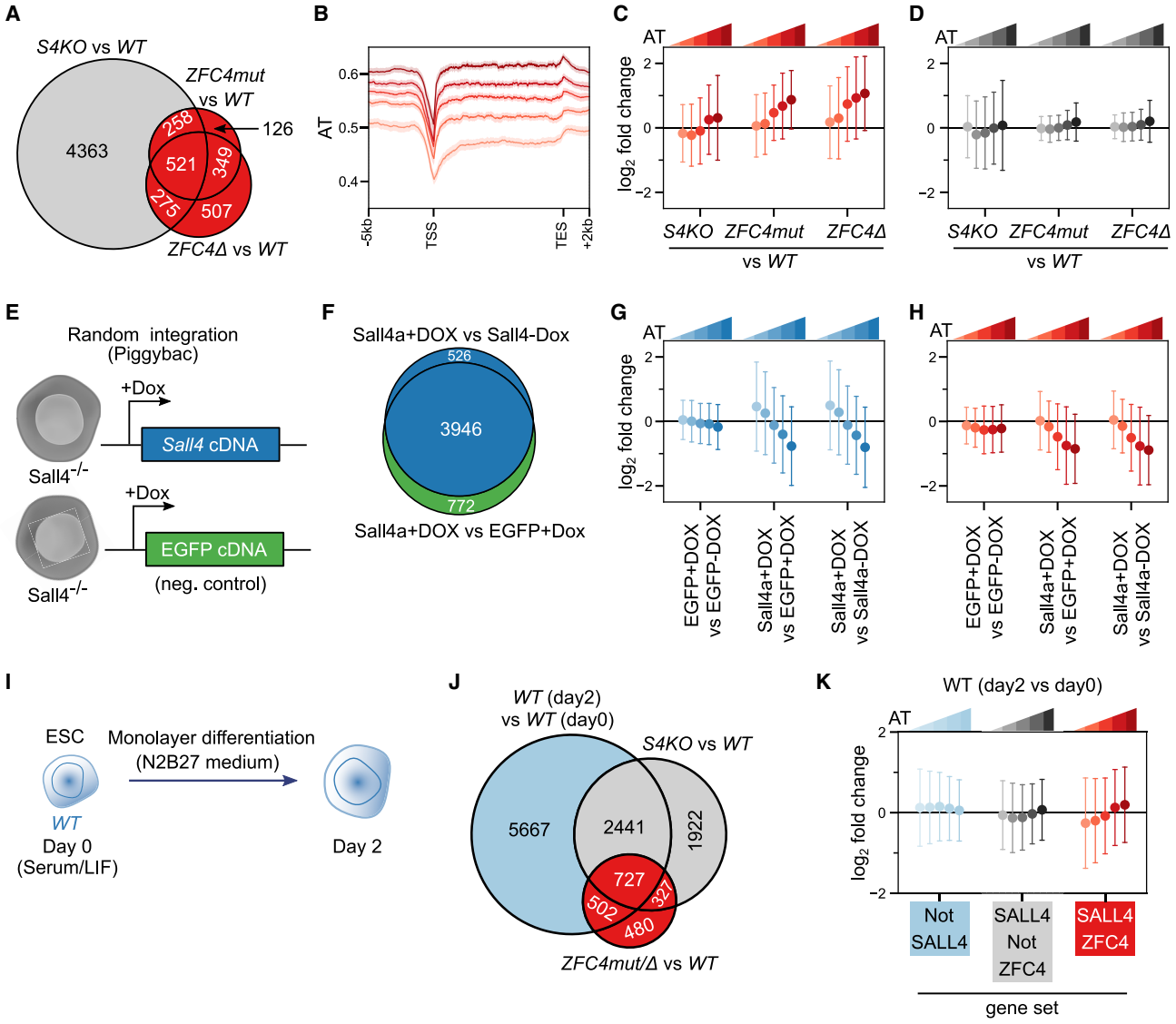


Figure 3. SALL4-mediated transcriptional regulation in relation to DNA base composition

(A) Venn diagram showing the overlap of differentially expressed genes detected by RNA-seq among *S4KO*, *ZFC4mut*, and *ZFC4Δ* ESCs. ZFC4-regulated genes are indicated in red and ZFC4-independent genes in gray.

(B) Profile plot showing the density of A/T nucleotides around the transcription unit of ZFC4-regulated genes divided into five equal categories according to AT content. TSS, transcription start site; TES, transcription end site.

(C and D) Correlation between gene mis-regulation (log2 fold change versus WT) and DNA base composition in *Sall4* mutant ESCs. ZFC4-regulated (C) and ZFC4-independent (D) genes were divided into five equal categories depending on their AT content.

(E) Diagram representing *S4KO* ESC lines carrying SALL4 or EGFP (control) expression constructs under control of a doxycycline-inducible promoter.

(F) Venn diagram showing the overlap of differentially expressed genes detected by RNA-seq following a 48-h doxycycline induction in the ESC lines presented in (E). SALL4-responsive genes are indicated in blue and EGFP-responsive genes in green.

(G and H) Correlation between SALL4-induced gene expression changes and DNA base composition. SALL4-responsive (G) and ZFC4-regulated (H) genes were divided into five equal categories depending on their AT content, and their relative expression levels were analyzed in the indicated ESC lines.

(I) Diagram showing the protocol used to characterize early differentiation of *WT* ESCs.

(J) Venn diagram showing the overlap between genes changing during early differentiation of *WT* cells (day 0 versus day 2) with genes de-regulated in *Sall4* mutant ESCs. Genes were divided into three categories: SALL4-independent genes (light blue), SALL4-dependent genes controlled by ZFC4 (red), and SALL4-dependent genes not controlled by ZFC4 (gray).

(K) Correlation between gene expression changes occurring during early differentiation and DNA base composition in *WT* cells. SALL4-independent genes (light blue), SALL4-dependent genes controlled by ZFC4 (red), and SALL4-dependent genes not controlled by ZFC4 (gray) were divided into five equal categories depending on their AT content, and their relative expression levels were analyzed at day 2 of differentiation.

See also Figure S3 and Tables S2 and S3.

Molecular Cell
Article

CellPress
OPEN ACCESS

significance of these relationships (Figures S3H and S3I). As a control, we applied the same analysis to the minority of genes whose expression was altered in response to EGFP induction (Figure 3F, green filling). In this case, there was no apparent relationship between fold change and base composition (Figures S3J and S3K), as confirmed by quantitative analysis (Figure S3L). Together, our results strongly suggest that SALL4 directly regulates gene expression in response to base composition via its zinc-finger cluster, ZFC4.

Interestingly, Gene Ontology (GO) analysis of ZFC4-regulated genes identified GO terms associated with neuronal differentiation, morphogenesis, gonad development, and kidney development (Table S3), all of which are adversely affected in *S4KO* mice and embryos (Sakaki-Yumoto et al., 2006; Böhm et al., 2008; Tahara et al., 2019; Koshiba-Takeuchi et al., 2006; Akiyama et al., 2015; Hobbs et al., 2012; Yamaguchi et al., 2015; Xu et al., 2017; Chan et al., 2017; Warren et al., 2007). This suggests the possibility that SALL4 plays an essential role in the transition between self-renewing ESCs and the differentiated state by preferentially suppressing the expression of AT-rich developmental genes, thus preventing premature loss of pluripotency. If so, AT-rich genes that are aberrantly upregulated in the absence of ZFC4 should be activated during the normal differentiation program of *WT* cells (Rao et al., 2010). To test this, we performed RNA-seq on *WT* ESCs following 2 days of monolayer differentiation (Figure 3I) (Aubert et al., 2002). Although they represent a small fraction of all transcriptional changes taking place at these stages, SALL4-regulated genes overlapped significantly with genes whose expression changes naturally between day 0 (ESCs) and day 2 of differentiation (Figure 3J). Importantly, ZFC4-regulated genes, but not other categories of genes, are upregulated at this early stage in proportion to AT richness (Figures 3K, S3M, and S3N). Thus, AT-rich genes that are repressed by SALL4 in ESCs are activated soon after the exit from pluripotency.

SALL4 ZFC4 is critical for neuronal differentiation and embryonic development

Previous work demonstrated that disruption of the *Sall4* gene leads to increased stem cell differentiation (Yuri et al., 2009; Miller et al., 2016). To test whether disrupting ZFC4 alone leads to phenotypic defects, we compared *ZFC4mut* and *S4KO* ESCs. Consistent with previous evidence showing that SALL4 is dispensable for the maintenance of pluripotency (Sakaki-Yumoto et al., 2006; Yuri et al., 2009; Tsubooka et al., 2009), both *S4KO* and *ZFC4mut* ESCs expressed normal levels of OCT4 (Figure S4A) and showed efficient self-renewal, with a modest decrease observed in *S4KO* ESCs (Figure S4B). Next, we used a monolayer differentiation assay, as described above, to assess the propensity of these cell lines to acquire a neuronal fate. After 5 days in N2B27 medium, ESCs lacking SALL4 or expressing a ZFC4 mutant protein generated many more TUJ1-positive cells compared to *WT* cells (Figure 4A). Further confirming increased neuronal differentiation, qRT-PCR analyses identified increased transcription of *Tuj1* (4- to 12-fold), *Ascl1* (3- to 6-fold), and *Nestin* (~2-fold) in *S4KO*, *ZFC4mut*, and *ZFC4Δ* ESCs at day 5 of differentiation (Figure 4B). By this assay, inactivation of ZFC4 phenocopies complete loss of SALL4 protein.

In order to observe the effects of ZFC4 mutation on embryonic development, we generated a *ZFC4mut* mouse line by blastocyst injection of heterozygous *Sall4*$^{ZFC4mut/WT}$ ESCs. F1 mice were crossed and their progeny analyzed at different stages of development. While *ZFC4mut* homozygous embryos were obtained at Mendelian ratios during early development, none survived until birth (Figures 5A and 5B). By embryonic day 10.5 (E10.5), homozygous embryos presented gross morphological abnormalities, which were not observed in controls (Figure 5C). Importantly, the ZFC4mut protein was expressed at levels similar to those seen in *WT* embryos (Figures S5A and S5B). Early embryonic mortality of *ZFC4* mutant mice is reminiscent of the phenotype observed in *S4KO* mice, although the latter die earlier in development, shortly after implantation (by E5.5–E6.5) (Elling et al., 2006; Sakaki-Yumoto et al., 2006). Taken together, our *in vitro* and *in vivo* experiments indicate that mutation of ZFC4 alone phenocopies important aspects of the *S4KO* phenotypes seen in both ESCs and embryos. It follows that this DNA-binding domain is a key contributor to SALL4 biological function.

C2H2 zinc-finger clusters ZFC1 and ZFC2 are dispensable for SALL4 function in ESCs

SALL4 contains two C2H2 zinc-finger clusters, ZFC1 and ZFC2, in addition to ZFC4. To determine their contribution to SALL4 function, we used CRISPR-Cas9 to delete the central segment of endogenous SALL4 protein, which contains the zinc-finger clusters ZFC1 and ZFC2, while leaving ZFC4 and the N-terminal domain intact (Figure 6A). ESCs homozygous for this *ZFC1-2Δ* knockin allele lack full-length SALL4, but, as expected, ZFC1-2Δ protein retained the ability to interact with SALL1 and NuRD components (Figure S6A). Immunostaining showed that ZFC1-2Δ resembled *WT* SALL4 by being enriched at heterochromatic foci, indicating that ZFC4 binding to this AT-rich DNA *in vivo* is unaffected by the internal deletion (Figure 6B). To characterize ZFC1-2Δ chromatin binding in more detail, we performed ChIP-seq (Figure S6B), as described above. In contrast to the dramatic effect of ZFC4 inactivation on SALL4 ChIP-seq peaks, ZFC1-2Δ occupancy of the genome closely resembled that of *WT* SALL4 (Figure S6C). In addition, both the average ChIP-seq signal (Figure 6C) and AT-rich profile (Figure 6D) of *WT* SALL4 peaks were preserved in *ZFC1-2Δ* cells. We conclude that ZFC1 and ZFC2 contribute minimally to the genome binding profile of SALL4, further supporting the view that ZFC4 is the primary determinant of DNA binding. Comparative RNA-seq analysis between *WT*, *ZFC4mut*, and *ZFC1-2Δ* ESCs revealed that SALL4 ZFC1-2Δ and ZFC4mut affect largely nonoverlapping sets of genes (Figure 6E). The effects of ZFC1-2Δ on transcription were independent of base composition, whereas ZFC4 regulated genes in proportion to their AT richness (Figure 6F, S6D, S6E and S6F). Finally, we examined the phenotypic consequences of ZFC1-2 deletion by assaying monolayer neuronal differentiation of our mutant ESCs. Unlike *S4KO* and *ZFC4mut* ESCs, *ZFC1-2Δ* cells did not show evidence of increased differentiation as assessed by TUJ1 immunofluorescence

CellPress OPEN ACCESS

Molecular Cell
Article

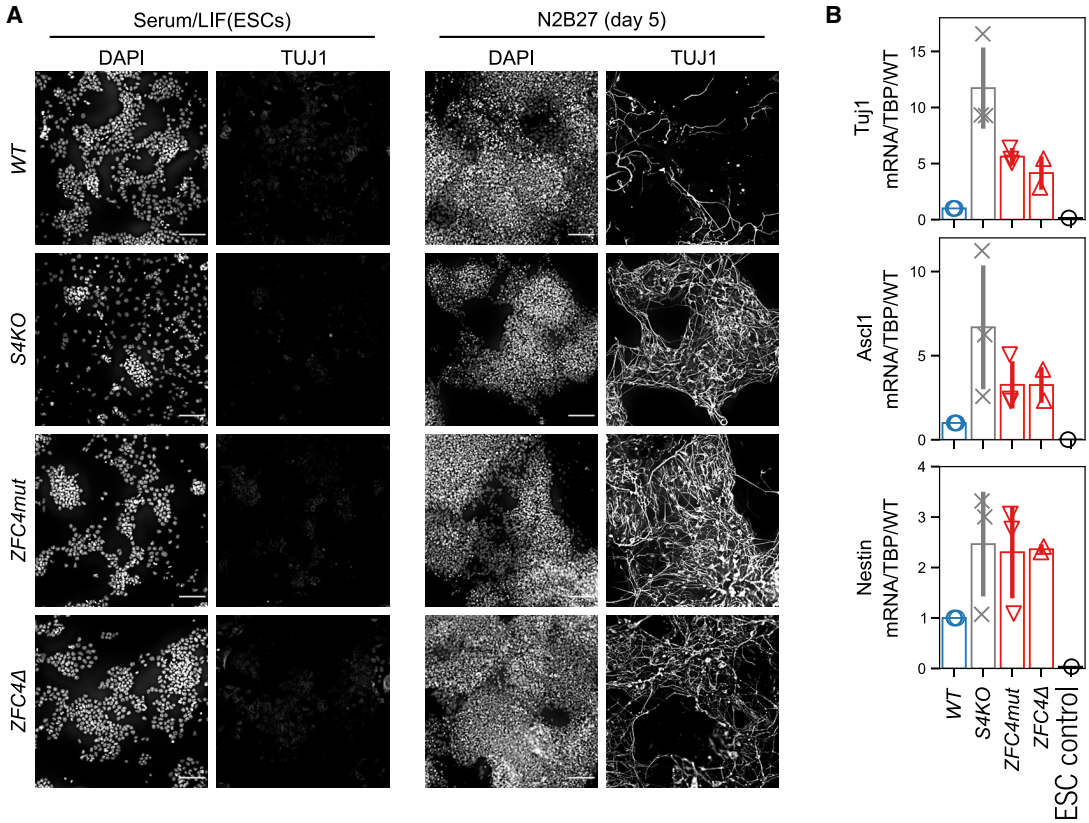


Figure 4. Phenotypic effects of SALL4 ZFC4 mutation on neuronal differentiation

(A) TUJ1 immunofluorescence in the indicated ESC lines cultured in serum/leukemia inhibitory factor (LIF) medium, and following differentiation for 5 days in N2B27 medium. DNA was stained with DAPI. Scale bars, 100 μm.

(B) qRT-PCR analysis of the neuronal markers Tuj1, Ascl1, and Nestin in the indicated cell lines following differentiation for 5 days in N2B27 medium. Transcript levels were normalized to TATA-binding protein (TBP) and expressed relative to *WT*. Data points indicate independent replicate experiments and error bars standard deviation.

See also Figure S4.

(Figure 6G) and RT-qPCR analysis of neuronal markers at day 5 of differentiation (Figure S6G).

To further characterize the differentiation defects of *Sall4* mutant ESCs, we performed an RNA-seq time-course experiment with *WT*, *S4KO*, *ZFC4mut*, and *ZFC1-2Δ* cell lines at days 0 (ESCs), 2, and 5 of the differentiation protocol (Figure 7A). In agreement with our previous base-composition analyses, absence of SALL4 or inactivation of ZFC4 weakened repression, leading to premature activation of ZFC4-regulated AT-rich genes at all differentiation time points (Figure S7A). In contrast, ZFC1/2-regulated genes showed no preferential upregulation during differentiation and no correlation with base composition in any of the cell lines (Figure S7B). Moreover, principal-component analysis (PCA) showed that *WT* and *ZFC1-2Δ* samples clustered together at all time points, while *S4KO* and *ZFC4mut* formed an independent cluster at days 2 and 5 (Figure S7C). Accordingly, differential expression analysis across our time series revealed few differences between *WT* and *ZFC1-2Δ*, while the transcriptomes of *S4KO* and *ZFC4mut* were significantly disturbed (Figure 7B). Emphasizing the similarity of *S4KO* and *ZFC4mut*, genes differentially regulated in these cell lines were highly correlated both at day 2 and day 5 of differentiation (Figures 7C and S7D). Also, genes associated with neuronal differentiation were upregulated in both cell lines, whereas expression of these genes in *ZFC1-2Δ* cells was unaffected (Figure 7D). We conclude that the characteristic premature differentiation phenotype associated with SALL4 deficiency is mimicked by inactivation of ZFC4, but not by a large deletion of the central domain that includes ZFC1 and ZFC2.

DISCUSSION

SALL4 targets a broad range of AT-rich motifs via the zinc-finger cluster ZFC4. While the ZFC4 domain has previously been implicated in binding to AT-rich repetitive DNA found in mouse major satellite (Yamashita et al., 2007), its biological significance was unknown. Our study demonstrates that ZFC4 is a key domain mediating SALL4 biological function in ESCs. Its inactivation drastically reduces peaks of SALL4 binding to the genome, suggesting that this domain plays a critical role in SALL4 targeting to chromatin. Disruption of the genomic binding pattern is

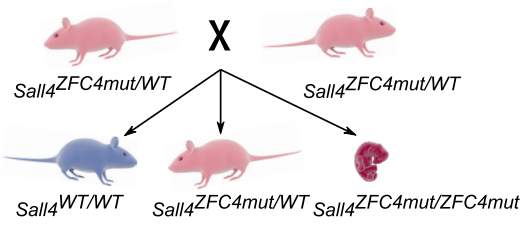
Molecular Cell
Article

CellPress
OPEN ACCESS

A

Strain	Stage	WT	Het	Hom	χ^2	p-value
ZFC4mut	E3.5	5	6	4	0.733	0.693
ZFC4mut	E9.5	7	9	12	5.357	0.0687
ZFC4mut	E10.5	3	11	4	1	0.6065
ZFC4mut	E11.5	5	15	10	1.667	0.4346
ZFC4mut	Pups	29	47	**0**	26.395	<0.0001

B



C

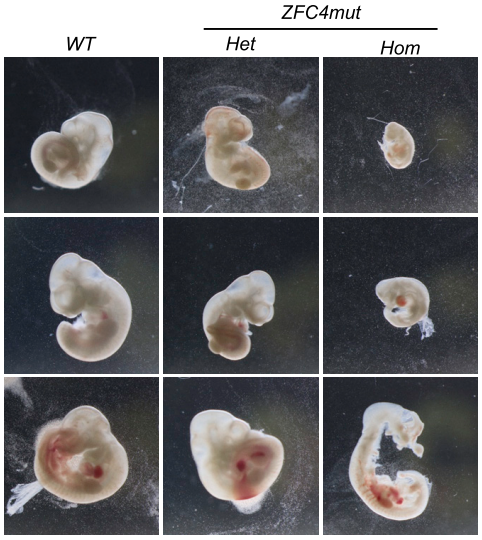


Figure 5. Mutation of SALL4 ZFC4 causes embryonic lethality

(A) Table showing the number of live pups and embryos at different stages of development and their associated genotype. Animals were crossed to obtain *ZFC4mut* heterozygous (*Het*), homozygous (*Hom*), or *WT* progeny.

(B) Diagram showing the results from crossing *ZFC4mut* heterozygote mice. *ZFC4mut* homozygous animals die during embryonic development.

(C) Representative images of *WT*, *ZFC4mut* heterozygous (*Het*), and homozygous (*Hom*) embryos at E10.5, taken at the same magnification.

See also Figure S5.

accompanied by mis-regulation of a subset of all SALL4-regulated genes, many of which are implicated in neuronal differentiation, which is the preferred fate of ESCs in culture. Accordingly, ESCs expressing SALL4 lacking a functional ZFC4 domain phenocopy *S4KO* ESCs by displaying precocious differentiation toward the neuronal lineage. Importantly, we found that this gene set is normally activated as *WT* ESCs commence differentiation in culture to give neurons, a process that coincides with downregulation of SALL4 protein (Rao et al., 2010). The importance of ZFC4 is further demonstrated by the embryonic lethal phenotype of *ZFC4mut* homozygotes.

Levels of ZFC4 mutant proteins are somewhat reduced in the mutant ESCs (50%–75% of WT), but we consider it unlikely that this contributes to the biased effect on expression of AT-rich genes. The changes in genome occupancy revealed by ChIP and immunostaining are much greater than 2-fold and are unlikely to be due to the relatively modest reduction in protein levels. In addition, the early embryonic lethality of the ZFC4 mutation is much more severe than that seen in mice heterozygous for the *Sall4 KO* allele, which can be viable and fertile despite having 50% less protein (Sakaki-Yumoto et al., 2006). Notably, the reduction in ZFC4mut protein seen in ESCs is not reproduced in mouse embryos, as mutant and WT proteins are present at indistinguishable levels, yet the phenotype is nevertheless extremely severe. On the other hand, loss of the only AT-binding domain in the protein offers an attractive explanation for this phenomenon. Human genetics provides further support for the central importance of ZFC4. Mutations in the *SALL4* gene cause Okihiro syndrome (Al-Baradie et al., 2002; Kohlhase et al., 2002; Terhal et al., 2006), with most patients carrying frameshift or nonsense mutations leading to deletion or severe truncation of the protein. The only reported disease-causing missense mutation (H888R) affects a zinc-coordinating histidine that is expected to specifically inactivate ZFC4 (Miertus et al., 2006), although this has not been tested experimentally.

Evidence regarding the functional significance of two other zinc-finger clusters, ZFC1 and ZFC2, is limited, although an affinity of ZFC1 for hydroxymethylcytosine has been reported (Xiong et al., 2016). Importantly, simultaneous deletion of ZFC1 and ZFC2 has a minimal effect on genome occupancy, gene expression, and propensity to differentiate of ESCs. Thus, the well-known role of SALL4 in stabilization of the pluripotent state appears to be largely attributable to the DNA-binding specificity of ZFC4. Our observations agree with previous studies using transfection assays that indicated that the naturally occurring isoform SALL4B, which closely resembles ZFC1-2Δ in lacking ZFC1 and ZFC2 and is expressed at much lower levels than the full-length SALL4A form, retains biological activity in pluripotent cells (Rao et al., 2010; Miller et al., 2016). Although these results suggest that these two C2H2 zinc-finger clusters are dispensable for SALL4 function in ESCs, we note that their sequence is highly conserved between fruit flies and humans. It is therefore likely that ZFC1 and ZFC2 are functional in other developmental contexts, such as limb development and/or gametogenesis.

At first sight, the correlation with base composition across the extended transcription unit contrasts with the relatively sharp peaks of SALL4 binding observed by ChIP-seq. In fact, it remains to be determined whether SALL4 acts at distance from AT-rich motifs in discrete regulatory elements or by binding broadly to AT-rich sequences dispersed through gene bodies. The latter would be challenging to detect by

CellPress OPEN ACCESS

Molecular Cell
Article

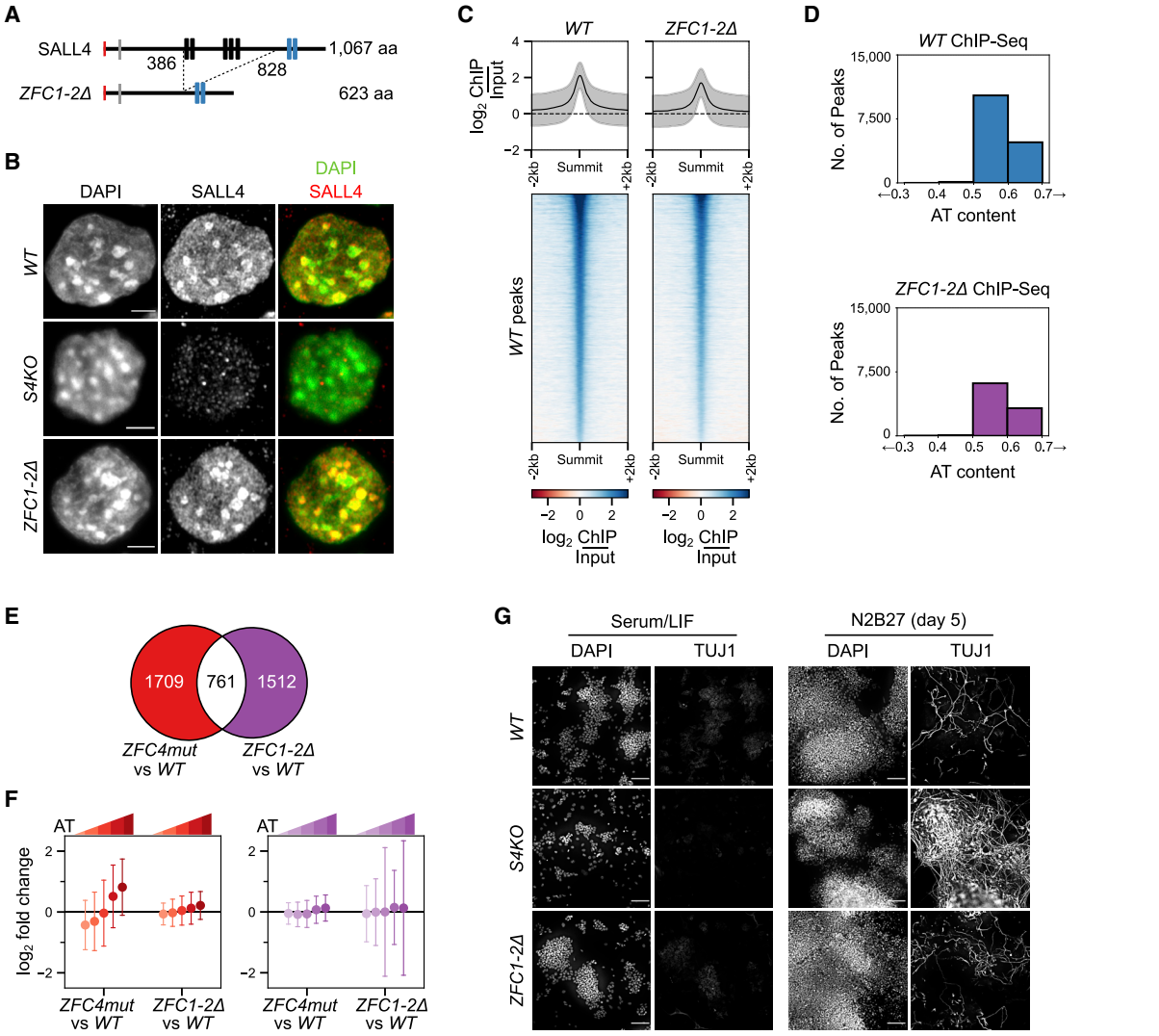


Figure 6. Effects of SALL4 ZFC1 and ZFC2 deletion in ESCs on chromatin binding, gene expression, and differentiation

(A) Diagram showing the in-frame deletion within the *Sall4* coding sequence, generated by CRISPR-Cas9.

(B) SALL4 ZFC1-2Δ localization determined by immunofluorescence in the indicated ESC lines. DNA was stained with DAPI, showing dense clusters of AT-rich pericentric chromatin. Scale bars, 3 μm.

(C) Heatmap and profile plot showing SALL4 ChIP-seq signal at SALL4 *WT* ChIP-seq peaks in the indicated cell lines.

(D) Analysis of the DNA base composition surrounding SALL4 ChIP-seq peaks (summit ±250 bp) in *WT* (blue) and *ZFC1-2Δ* (purple) ESCs.

(E) Venn diagram showing the overlap of differentially expressed genes detected by RNA-seq between *ZFC4mut* and *ZFC1-2Δ* ESCs. ZFC4-regulated genes are indicated in red and ZFC1/2-regulated genes in purple.

(F) Correlation between gene mis-regulation (log2 fold change *versus WT*) and DNA base composition in *Sall4* mutant ESCs. ZFC4-regulated (red) and ZFC1/2-regulated (purple) genes were divided into five equal categories depending on their AT content.

(G) TUJ1 immunofluorescence in the indicated ESC lines cultured in serum/LIF medium and following differentiation for 5 days in N2B27 medium. DNA was stained with DAPI. Scale bars, 100 μm.

See also Figure S6 and Table S2.

ChIP due to the high abundance of AT-rich motifs throughout the genome (potentially in excess of 10 million target sites) in contrast with the low abundance of SALL4 protein in ESCs (2,000–3,000 copies per cell) (Zhang et al., 2017). As a result of this discrepancy, percent occupancy of any one target site is likely to be extremely low. Further work is required to distinguish the effects on gene expression of dispersed versus focal SALL4 binding. An obvious potential mediator of repression by SALL4 is the NuRD co-repressor complex, which has long been known to associate with the N terminus of SALL4 (Lauberth and Rauchman, 2006). The role of NuRD recruitment for SALL4 function has been questioned, however (Miller et al.,

Molecular Cell
Article

CellPress
OPEN ACCESS

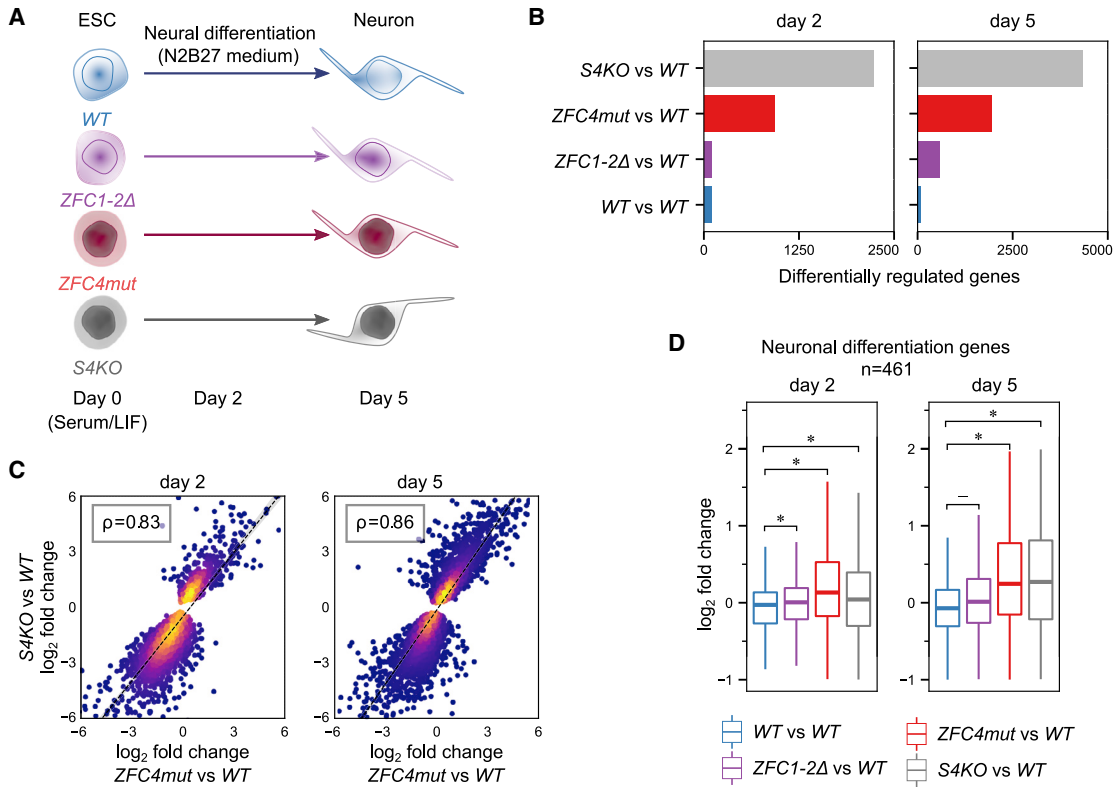


Figure 7. Transcriptional effects of SALL4 zinc-finger cluster mutations during neuronal differentiation

(A) Diagram of the RNA-seq time-course experiment comparing the differentiation potential of *WT* and *Sall4* mutant ESCs.

(B) Differential gene expression analysis between *WT* and *Sall4* mutant cell lines during neuronal differentiation at days 2 and 5 (adjusted p value < 0.05). Additional *WT* replicates were used as a control (*WT* versus *WT*).

(C) Scatterplot showing the relative expression levels of genes deregulated in differentiating *S4KO* cells (see B, gray bars) correlating with their expression in *ZFC4mut* cells at days 2 and 5 of differentiation.

(D) Relative expression levels (log2 fold change versus *WT*) of genes associated with the GO term "positive regulation of neuron differentiation" (GO:0045666) in *Sall4* mutant cell lines at days 2 and 5 of differentiation. Additional *WT* replicates were used as a control (*WT* versus *WT*). Stars indicate statistical significance calculated from bootstrapped two-sided t tests (p value < 0.05).

See also Figure S7 and Table S2.

2016). Another poorly understood aspect of SALL4 biochemistry is its interaction with other members of the SALL family (Kiefer et al., 2003; Sweetman et al., 2003). Notably, our screen for AT-binding proteins also identified SALL1 and SALL3, which both interact with SALL4 and might contribute to sensing AT content via their closely similar ZFC4 domains. Given the importance of SALL4 in development and disease, these issues deserve further investigation.

Our results demonstrate that cell-type-specific genes residing within AT-rich domains are susceptible to repression by the transcriptional repressor, SALL4, thereby preventing differentiation. Vertebrate genomes are on average relatively AT rich (60% A/T), and therefore, the short A/T motifs to which it binds occur throughout the genome with frequencies that vary probabilistically according to local base composition. As base composition is a constant feature of the genome, regulation is achieved by varying the availability of the base composition reader itself. Accordingly, as cells enter differentiation, expression of SALL4 drops (Rao et al., 2010), raising the possibility that differentiation is triggered by loss of SALL4-mediated inhibition of key developmental genes. Global regulation of this kind confers the ability to modulate expression of multi-gene blocks using relatively few base composition readers and is potentially more economical than controlling each gene by a separate mechanism. Our finding that this relatively simple system may underlie large-scale switching of gene expression programs indicates that base compositional domains are not merely a biologically irrelevant byproduct of genome evolution but constitute a signal that is advantageous to the organism.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact

CellPress
OPEN ACCESS

Molecular Cell
Article

- Materials availability
- Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - *In vivo* animal studies (mouse)
  - *In vitro* cell culture studies (mouse)
  - Genetic manipulation of ESCs
- METHOD DETAILS
  - DNA pulldown and mass spectrometry
  - Immunoprecipitation
  - Western blot
  - Immunofluorescence
  - RT-qPCR
  - HT-Selex
  - ChIP-seq
  - ATAC-seq
  - RNA-seq
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - HT-SELEX analysis
  - RNA-seq analysis
  - ChIP-seq analysis
  - ATAC-seq analysis
  - Quantification of AT effect

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2020.11.046.

ACKNOWLEDGMENTS

We thank Sara Giuliani for providing critical feedback on the manuscript, Dina De Sousa and Michal Prendecki for technical assistance, Shaun Webb for bioinformatics advice, David Kelly and the Centre Optical Instrumentation Laboratory facility for microscopy support, Martin Waterfall for support with flow cytometry, the Edinburgh Protein Production Facility for access to protein purification facilities, and Vladimir Benes (EMBL GeneCore facility, Germany) for support with high-throughput sequencing. We are grateful to Riuchi Nishinakamura (Kumamoto University) and Brian Hendrich (Cambridge University) for sharing *S4KO* ESCs and SALL4 expression plasmids. The Edinburgh Protein Production Facility and the Centre Optical Instrumentation Laboratory are funded by core grants (092076 and 203149) to the Wellcome Centre for Cell Biology at the University of Edinburgh. This work was primarily funded by a European Research Council advanced grant EC 694295 (Gen-Epix) and Wellcome Investigator Award 107930 to A.B., who is also a member of the Simons Initiative for the Developing Brain. K.S.S. was supported by a Sir Henry Wellcome Fellowship (grant 101489/Z/13/Z). T.Q. received an EU Marie Curie Fellowship. A.C. is supported by Wellcome Senior Fellowship 200898. The M.V. lab is part of the Oncode Institute, which is funded by the Dutch Cancer Society. C.G.S. is supported by NWO-VENI grant 722.016.003.

AUTHOR CONTRIBUTIONS

Conceptualization, A.B., T.Q., R.P., and K.C.; Methodology, R.P., K.C., T.Q., J.C.-W., C.G.S., M.V., and J.S.; Software, K.C.; Formal Analysis, K.C.; Investigation, R.P., K.C., T.Q., K.S.-S., G.A., B.A.-H., H.Y.L., A.C., J.C.-W., C.G.S., and J.S.; Writing – Original Draft, R.P., K.C., and A.B.; Writing – Review & Editing, R.P., K.C., and A.B.; Supervision, A.B.; Funding acquisition, A.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 29, 2020
Revised: October 23, 2020
Accepted: November 25, 2020
Published: January 5, 2021

REFERENCES

Akiyama, R., Kawakami, H., Wong, J., Oishi, I., Nishinakamura, R., and Kawakami, Y. (2015). Sall4-Gli3 system in early limb progenitors is essential for the development of limb skeletal elements. Proc. Natl. Acad. Sci. USA *112*, 5075–5080.

Al-Baradie, R., Yamada, K., St Hilaire, C., Chan, W.-M., Andrews, C., McIntosh, N., Nakano, M., Martonyi, E.J., Raymond, W.R., Okumura, S., et al. (2002). Duane radial ray syndrome (Okihiro syndrome) maps to 20q13 and results from mutations in SALL4, a new member of the SAL family. Am. J. Hum. Genet. *71*, 1195–1199.

Aravind, L., and Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res. *26*, 4413–4421.

Arhondakis, S., Auletta, F., and Bernardi, G. (2011). Isochores and the regulation of gene expression in the human genome. Genome Biol. Evol. *3*, 1080–1089.

Aubert, J., Dunstan, H., Chambers, I., and Smith, A. (2002). Functional gene screening in embryonic stem cells implicates Wnt antagonism in neural differentiation. Nat. Biotechnol. *20*, 1240–1245.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science *228*, 953–958.

Bickmore, W.A., and Sumner, A.T. (1989). Mammalian chromosome banding: an expression of genome organization. Trends Genet. *5*, 144–148.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. Mol. Cell *38*, 179–190.

Böhm, J., Buck, A., Borozdin, W., Mannan, A.U., Matysiak-Scholze, U., Adham, I., Schulz-Schaeffer, W., Floss, T., Wurst, W., Kohlhase, J., and Barrionuevo, F. (2008). Sall1, sall2, and sall4 are required for neural tube closure in mice. Am. J. Pathol. *173*, 1455–1463.

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

Cerda, M.C., Berríos, S., Fernández-Donoso, R., Garagna, S., and Redi, C. (1999). Organisation of complex nuclear domains in somatic mouse cells. Biol. Cell *91*, 55–65.

Chan, A.-L., La, H.M., Legrand, J.M.D., Mäkelä, J.-A., Eichenlaub, M., De Seram, M., Ramialison, M., and Hobbs, R.M. (2017). Germline stem cell activity is sustained by SALL4-dependent silencing of distinct tumor suppressor genes. Stem Cell Reports *9*, 956–971.

Cholewa-Waclaw, J., Shah, R., Webb, S., Chhatbar, K., Ramsahoye, B., Pusch, O., Yu, M., Greulich, P., Waclaw, B., and Bird, A.P. (2019). Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. Proc. Natl. Acad. Sci. USA *116*, 14995–15000.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442–459.e20.

Costantini, M., Cammarano, R., and Bernardi, G. (2009). The evolution of isochore patterns in vertebrate genomes. BMC Genomics *10*, 146.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Dey, K.K., Xie, D., and Stephens, M. (2018). A new sequence logo plot to highlight enrichment and depletion. BMC Bioinformatics *19*, 473.

Molecular Cell
Article

CellPress
OPEN ACCESS

Donovan, K.A., An, J., Nowak, R.P., Yuan, J.C., Fink, E.C., Berry, B.C., Ebert, B.L., and Fischer, E.S. (2018). Thalidomide promotes degradation of SALL4, a transcription factor implicated in Duane radial ray syndrome. eLife *7*, e38430.

Duret, L., Eyre-Walker, A., and Galtier, N. (2006). A new perspective on isochore evolution. Gene *385*, 71–74.

Elling, U., Klasen, C., Eisenberger, T., Anlag, K., and Treier, M. (2006). Murine inner cell mass-derived lineages depend on Sall4 function. Proc. Natl. Acad. Sci. USA *103*, 16319–16324.

Eyre-Walker, A., and Hurst, L.D. (2001). The evolution of isochores. Nat. Rev. Genet. *2*, 549–555.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. eLife *1*, e00205.

Filarsky, M., Zillner, K., Araya, I., Villar-Garea, A., Merkl, R., Längst, G., and Németh, A. (2015). The extended AT-hook is a novel RNA binding motif. RNA Biol. *12*, 864–876.

Guenatri, M., Bailly, D., Maison, C., and Almouzni, G. (2004). Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. J. Cell Biol. *166*, 493–505.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nat. Cell Biol. *15*, 373–384.

Hiratani, I., Takebayashi, S., Lu, J., and Gilbert, D.M. (2009). Replication timing and transcriptional control: beyond cause and effect–part II. Curr. Opin. Genet. Dev. *19*, 142–149.

Hobbs, R.M., Fagoonee, S., Papa, A., Webster, K., Altruda, F., Nishinakamura, R., Chai, L., and Pandolfi, P.P. (2012). Functional antagonism between Sall4 and Plzf defines germline progenitors. Cell Stem Cell *10*, 284–298.

Holmquist, G.P. (1989). Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. *28*, 469–486.

Holmquist, G., Gray, M., Porter, T., and Jordan, J. (1982). Characterization of Giemsa dark- and light-band DNA. Cell *31*, 121–129.

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. Nature *326*, 292–295.

Jabbari, K., and Bernardi, G. (2017). An isochore framework underlies chromatin architecture. PLoS ONE *12*, e0168023.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol. *12*, 918–922.

Kiefer, S.M., Ohlemiller, K.K., Yang, J., McDill, B.W., Kohlhase, J., and Rauchman, M. (2003). Expression of a truncated Sall1 transcriptional repressor is responsible for Townes-Brocks syndrome birth defects. Hum. Mol. Genet. *12*, 2221–2227.

Kohlhase, J., Heinrich, M., Schubert, L., Liebers, M., Kispert, A., Laccone, F., Turnpenny, P., Winter, R.M., and Reardon, W. (2002). Okihiro syndrome is caused by SALL4 mutations. Hum. Mol. Genet. *11*, 2979–2987.

Koshiba-Takeuchi, K., Takeuchi, J.K., Arruda, E.P., Kathiriya, I.S., Mo, R., Hui, C.C., Srivastava, D., and Bruneau, B.G. (2006). Cooperative and antagonistic interactions between Sall4 and Tbx5 pattern the mouse limb and heart. Nat. Genet. *38*, 175–183.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Lauberth, S.M., and Rauchman, M. (2006). A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex. J. Biol. Chem. *281*, 23922–23931.

Lee, J.-H., Voo, K.S., and Skalnik, D.G. (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. J. Biol. Chem. *276*, 44669–44676.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997 https://arxiv.org/abs/1303.3997.

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*, 550.

Lu, J., Jeong, H.W., Kong, N., Yang, Y., Carroll, J., Luo, H.R., Silberstein, L.E., Yupoma, and Chai, L. (2009). Stem cell factor SALL4 represses the transcriptions of PTEN and SALL1 through an epigenetic repressor complex. PLoS ONE *4*, e5577.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Matsuda, Y., and Chapman, V.M. (1991). In situ analysis of centromeric satellite DNA segregating in Mus species crosses. Mamm. Genome *1*, 71–77.

Matyskiela, M.E., Couto, S., Zheng, X., Lu, G., Hui, J., Stamp, K., Drew, C., Ren, Y., Wang, M., Carpenter, A., et al. (2018). SALL4 mediates teratogenicity as a thalidomide-dependent cereblon substrate. Nat. Chem. Biol. *14*, 981–987.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Miertus, J., Borozdin, W., Frecer, V., Tonini, G., Bertok, S., Amoroso, A., Miertus, S., and Kohlhase, J. (2006). A *SALL4* zinc finger missense mutation predicted to result in increased DNA binding affinity is associated with cranial midline defects and mild features of Okihiro syndrome. Hum. Genet. *119*, 154–161.

Miller, A., Ralser, M., Kloet, S.L., Loos, R., Nishinakamura, R., Bertone, P., Vermeulen, M., and Hendrich, B. (2016). Sall4 controls differentiation of pluripotent cells independently of the nucleosome remodelling and deacetylation (NuRD) complex. Development *143*, 3074–3084.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *4*, e04837.

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.

Patsialou, A., Wilsker, D., and Moran, E. (2005). DNA-binding properties of ARID family proteins. Nucleic Acids Res. *33*, 66–80.

Quante, T., and Bird, A. (2016). Do short, frequent DNA sequence motifs mould the epigenome? Nat. Rev. Mol. Cell Biol. *17*, 257–262.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187-91.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Rao, S., Zhen, S., Roumiantsev, S., McDonald, L.T., Yuan, G.-C., and Orkin, S.H. (2010). Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. Mol. Cell. Biol. *30*, 5364–5380.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.

Sakaki-Yumoto, M., Kobayashi, C., Sato, A., Fujimura, S., Matsumoto, Y., Takasato, M., Kodama, T., Aburatani, H., Asashima, M., Yoshida, N., and Nishinakamura, R. (2006). The murine homolog of SALL4, a causative gene in Okihiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with Sall1 in anorectal, heart, brain and kidney development. Development *133*, 3005–3013.

CellPress
OPEN ACCESS

Molecular Cell
Article

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W.T.C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., et al. (2013a). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell *152*, 1146–1159.

Spruijt, C.G., Baymaz, H.I., and Vermeulen, M. (2013b). Identifying specific protein–dna interactions using SILAC-based quantitative proteomics. In Gene Regulation: Methods and Protocols, M. Bina, ed. (Humana Press), pp. 137–157.

Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G., and Pombo, A. (2007). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat. Cell Biol. *9*, 1428–1435.

Sweetman, D., and Münsterberg, A. (2006). The vertebrate spalt genes in development and disease. Dev. Biol. *293*, 285–293.

Sweetman, D., Smith, T., Farrell, E.R., Chantry, A., and Münsterberg, A. (2003). The conserved glutamine-rich region of chick csal1 and csal3 mediates protein interactions with other spalt family members. Implications for Townes-Brocks syndrome. J. Biol. Chem. *278*, 6560–6566.

Tahara, N., Kawakami, H., Chen, K.Q., Anderson, A., Yamashita Peterson, M., Gong, W., Shah, P., Hayashi, S., Nishinakamura, R., Nakagawa, Y., et al. (2019). *Sall4* regulates neuromesodermal progenitors and their descendants during body elongation in mouse embryos. Development *146*, dev177659.

Tanimura, N., Saito, M., Ebisuya, M., Nishida, E., and Ishikawa, F. (2013). Stemness-related factor Sall4 interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. J. Biol. Chem. *288*, 5027–5038.

Terhal, P., Rösler, B., and Kohlhase, J. (2006). A family with features overlapping Okihiro syndrome, hemifacial microsomia and isolated Duane anomaly caused by a novel SALL4 mutation. Am. J. Med. Genet. A. *140*, 222–226.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Tong, J.K., Hassig, C.A., Schnitzler, G.R., Kingston, R.E., and Schreiber, S.L. (1998). Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. Nature *395*, 917–921.

Tsubooka, N., Ichisaka, T., Okita, K., Takahashi, K., Nakagawa, M., and Yamanaka, S. (2009). Roles of Sall4 in the generation of pluripotent stem cells from blastocysts and fibroblasts. Genes Cells *14*, 683–694.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Uhlen, M., Bandrowski, A., Carr, S., Edwards, A., Ellenberg, J., Lundberg, E., Rimm, D.L., Rodriguez, H., Hiltke, T., Snyder, M., and Yamamoto, T. (2016). A proposal for validation of antibodies. Nat. Methods *13*, 823–827.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., and Skalnik, D.G. (2000). Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. Mol. Cell. Biol. *20*, 2108–2121.

Wade, P.A., Jones, P.L., Vermaak, D., and Wolffe, A.P. (1998). A multiple subunit Mi-2 histone deacetylase from Xenopus laevis cofractionates with an associated Snf2 superfamily ATPase. Curr. Biol. *8*, 843–846.

Warren, M., Wang, W., Spiden, S., Chen-Murchie, D., Tannahill, D., Steel, K.P., and Bradley, A. (2007). A Sall4 mutant mouse model useful for studying the role of Sall4 in early embryonic development and organogenesis. Genesis *45*, 51–58.

Wu, X., Johansen, J.V., and Helin, K. (2013). Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. Mol. Cell *49*, 1134–1146.

Xiong, J., Zhang, Z., Chen, J., Huang, H., Xu, Y., Ding, X., Zheng, Y., Nishinakamura, R., Xu, G.-L., Wang, H., et al. (2016). Cooperative action between SALL4A and TET proteins in stepwise oxidation of 5-methylcytosine. Mol. Cell *64*, 913–925.

Xu, K., Chen, X., Yang, H., Xu, Y., He, Y., Wang, C., Huang, H., Liu, B., Liu, W., Li, J., et al. (2017). Maternal Sall4 is indispensable for epigenetic maturation of mouse oocytes. J. Biol. Chem. *292*, 1798–1807.

Xue, Y., Wong, J., Moreno, G.T., Young, M.K., Côté, J., and Wang, W. (1998). NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. Mol. Cell *2*, 851–861.

Yamaguchi, Y.L., Tanaka, S.S., Kumagai, M., Fujimoto, Y., Terabayashi, T., Matsui, Y., and Nishinakamura, R. (2015). Sall4 is essential for mouse primordial germ cell specification by suppressing somatic cell program genes. Stem Cells *33*, 289–300.

Yamashita, K., Sato, A., Asashima, M., Wang, P.-C., and Nishinakamura, R. (2007). Mouse homolog of SALL1, a causative gene for Townes-Brocks syndrome, binds to A/T-rich sequences in pericentric heterochromatin via its C-terminal zinc finger domains. Genes Cells *12*, 171–182.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Yuri, S., Fujimura, S., Nimura, K., Takeda, N., Toyooka, Y., Fujimura, Y., Aburatani, H., Ura, K., Koseki, H., Niwa, H., and Nishinakamura, R. (2009). Sall4 is essential for stabilization, but not for pluripotency, of embryonic stem cells by repressing aberrant trophectoderm gene expression. Stem Cells *27*, 796–805.

Yusa, K., Zhou, L., Li, M.A., Bradley, A., and Craig, N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. Proc. Natl. Acad. Sci. USA *108*, 1531–1536.

Zhang, Y., LeRoy, G., Seelig, H.-P., Lane, W.S., and Reinberg, D. (1998). The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. Cell *95*, 279–289.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, X., Smits, A.H., van Tilburg, G.B.A., Jansen, P.W.T.C., Makowski, M.M., Ovaa, H., and Vermeulen, M. (2017). An interaction landscape of ubiquitin signaling. Mol. Cell *65*, 941–955.e8.

Molecular Cell
Article

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
SALL4	Abcam	ab29112; RRID: AB_777810
SALL4	Santa Cruz Biotechnology	sc-101147; RRID: AB_1129262
CHD4	Abcam	ab70469; RRID: AB_2229454
MTA1	Santa Cruz Biotechnology	sc-9446; RRID:AB_649545
MTA1	Santa Cruz Biotechnology	sc-373765; RRID: AB_10917039
SALL1	Abcam	ab41974; RRID: AB_777807
GATAD2A	Abcam	ab87663; RRID: AB_1952305
MBD3	Bethyl	A302-528A; RRID: AB_1998980
HDAC1	Santa Cruz Biotechnology	sc-6298; RRID: AB_2279712
HDAC1	Santa Cruz Biotechnology	sc-81598; RRID: AB_2118083
OCT4	Abcam	ab19857; RRID: AB_445175
TUJ1	BioLegend	801201; RRID: AB_2313773
Chemicals, peptides, and recombinant proteins		
Recombinant ZFC4 (SALL4 residues 859-1028)	This paper	N/A
Recombinant ZFC1 (SALL4 residues 320-486)	This paper	N/A
Recombinant ZFC2 (SALL4 residues 551-662)	This paper	N/A
Critical commercial assays		
KAPA RNA Hyperprep Kit with RiboErase (RNA-seq library preparation)	Roche	08098131702
KAPA Hyperprep Kit (ChIP-seq library preparation)	Roche	07962347001
Deposited data		
Raw and processed data used to generate the Figures	This paper	Mendeley Data: https://doi.org/10.17632/rwzttj9pn2.1
RNA-seq of WT, S4KO, ZFC4mut and ZFC4Δ ESCs	This paper	Array Express: E-MTAB-7343
RNA-seq of WT, S4KO, ZFC4mut and ZFC1-2Δ ESCs	This paper	Array Express: E-MTAB-7655
ChIP-seq of anti-SALL4 in WT, S4KO, ZFC4mut and ZFC1-2Δ ESCs	This paper	Array Express: E-MTAB-9197
Time-course (day 0, 2 and 5) RNA-seq of WT, S4KO, ZFC4mut and ZFC1-2Δ ESCs	This paper	Array Express: E-MTAB-9198
RNA-seq of S4KO cells integrated with Sall4 cDNA or EGFP cDNA with a doxycycline inducible promoter	This paper	Array Express: E-MTAB-9202
HT-SELEX of recombinant C2H2 zinc finger domains of SALL4	This paper	Array Express: E-MTAB-9236
ATAC-seq in WT ESCs	This paper	Array Express: E-MTAB-9245
ChIP-seq of H3K4me1 and H3K27ac in WT ESCs	Chronis et al., 2017	Gene Expression Omnibus: GSE90893

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: cell lines		
Mouse: *Wild-type* embryonic stem cell line; E14Ju09	Hooper et al., 1987	N/A
Mouse: *Sall4* knockout embryonic stem cell line	Provided by Brian Hendrich laboratory (Miller et al., 2016)	N/A
Mouse: *Sall4 ZFC4mut* embryonic stem cell line	This paper	N/A
Mouse: *Sall4 ZFC4Δ* embryonic stem cell line	This paper	N/A
Mouse: *Sall4 ZFC1-2Δ* embryonic stem cell line	This paper	N/A
Mouse: *Sall4* knockout embryonic stem cell line with doxycycline-inducible Sall4 construct	This paper	N/A
Mouse: *Sall4* knockout embryonic stem cell line with doxycycline-inducible EGFP construct	This paper	N/A
Experimental models: organisms/strains		
Mouse: *Sall4 ZFC4mut*	This paper	N/A
Oligonucleotides		
DNA pulldown oligonucleotides	This paper; see Table S1	N/A
CRISPR/Cas9 oligonucleotides and genotyping primers	This paper; see Table S1	N/A
RT-qPCR primers	This paper; see Table S1	N/A
HT-SELEX oligonucleotides	This paper; see Table S1	N/A
Recombinant DNA		
Cas9/gRNA expression plasmid	Ran et al., 2013	Addgene PX330
PiggyBac vector with doxycycline-inducible construct and Hygromycin resistance cassette	This paper	N/A
PiggyBac vector with constitutively expressed Tet-On 3G and Zeocin resistance cassette	This paper	N/A
Hyperactive PiggyBac transposase expression plasmid	Yusa et al., 2011	N/A
Software and Algorithms		
Logolas	Dey et al., 2018	N/A
sailfish v0.9.2	Patro et al., 2014	N/A
DESeq2 v1.28.0	Love et al., 2014	N/A
bedtools nuc	Quinlan and Hall, 2010	N/A
deepTools	Ramírez et al., 2014	N/A
clusterProfiler	Yu et al., 2012	N/A
Trimmomatic v0.33	Bolger et al., 2014	N/A
bwa-mem v0.7.17	Li, 2013)	N/A
Picard toolkit	http://broadinstitute.github.io/picard/	N/A
computeGCBias	Benjamini and Speed, 2012	N/A
MACS v2.1.2	Zhang et al., 2008	N/A
MEME-ChIP v5.1.0	Machanick and Bailey, 2011	N/A
Bioinformatics analysis – command line arguments	This paper (Methods S1)	N/A

Molecular Cell
Article

CellPress
OPEN ACCESS

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Adrian Bird (a.bird@ed.ac.uk).

Materials availability

Reagents generated in this study are available upon request from the Lead Contact.

Data and code availability

Raw and processed high-throughput sequencing data was deposited on Array Express, as described in the Table below. Python scripts and source code used for bioinformatic analyses, raw western blot and microscopy images, as well as other types of unprocessed and processed data used to generate the figures are available on Mendeley Data (https://doi.org/10.17632/rwzttj9pn2.1).

For H3K4me1 and H3K27ac ChIP-seq in ESCs, previously published data were obtained from GEO (accession number: GSE90893).

Accession	Description
E-MTAB-7343	RNA-seq of *WT*, *S4KO*, *ZFC4mut* and *ZFC4Δ* ESCs
E-MTAB-7655	RNA-seq of *WT*, *S4KO*, *ZFC4mut* and *ZFC1-2Δ* ESCs
E-MTAB-9197	SALL4 ChIP-seq in *WT*, *S4KO*, *ZFC4mut* and *ZFC1-2Δ* ESCs
E-MTAB-9198	Time course RNA-seq during differentiation (day 0, 2 and 5) of *WT*, *S4KO*, *ZFC4mut* and *ZFC1-2Δ* ESCs
E-MTAB-9202	RNA-seq of *S4KO* cells carrying Sall4 cDNA or EGFP cDNA under a doxycycline inducible promoter
E-MTAB-9236	HT-SELEX of recombinant C2H2 zinc-finger domains of SALL4
E-MTAB-9245	ATAC-seq in *WT* ESCs

EXPERIMENTAL MODEL AND SUBJECT DETAILS

In vivo animal studies (mouse)

The *Sall4 ZFC4mut* mouse line was generated by injection of CRISPR/Cas9-targeted heterozygous ESCs (see section above) into mouse blastocysts using standard methods. Resultant male and female chimeras were crossed with C57BL/6J *wild-type* animals at 7 weeks and coat color was used to identify germline offspring. Transmission of the targeted allele was confirmed by PCR (see primers) and Sanger sequencing. Heterozygotes identified from these crosses were inter-crossed to generate homozygotes. Animals were routinely genotyped by PCR combined with restriction fragment length polymorphism (RFLP) analysis using *HaeII* (restriction site introduced within *ZFC4mut* allele). Heterozygosity had no discernible phenotype in either sex at any age. The sex of homozygous blastocysts and embryos was not determined post-mortem.

All mice used in this study were bred and maintained at the University of Edinburgh animal facilities under standard conditions, and procedures were carried out by staff licensed by the UK Home Office and in accordance with the Animal and Scientific Procedures Act 1986 following initial approval by a local Animal Welfare and Ethical Review Board. All mice were housed within a SPF facility. They were maintained on a 12h light/dark cycle and given *ad libitum* access to food and water. They were housed in open top cages with wood chippings, tissue bedding and additional environmental enrichment in groups of up to ten animals. Mutant mice were caged with their wild-type littermates.

In vitro cell culture studies (mouse)

Cell culture conditions

E14Ju09, a clonal cell line derived from E14Tg2a ESCs (Hooper et al., 1987), was used as a *wild-type* cell line in this study. *Sall4 ZFC4mut*, *ZFC4Δ*, and *ZFC1-2Δ* ESC lines were derived from E14Ju09 ESCs using CRISPR/Cas9, as indicated below. *Sall4*

CellPress
OPEN ACCESS

Molecular Cell
Article

knockout ESCs were kindly provided by Brian Hendrich (Cambridge University) with agreement of Riuchi Nishinakamura (Kumamoto University) (Miller et al., 2016). SALL4 and EGFP doxycycline-inducible ESC lines were derived from *Sall4* knockout ESCs using the PiggyBac (PB) transposon system, as indicated below.

All ESC lines were incubated at 37°C and 5% CO_2 in gelatin-coated dishes containing Glasgow minimum essential medium (GMEM; GIBCO ref. 11710035) supplemented with 15% fetal bovine serum (batch tested), 1x L-glutamine (GIBCO ref. 25030024), 1x MEM non-essential amino acids (GIBCO ref.11140035), 1mM sodium pyruvate (GIBCO ref. 11360039), 0.1mM 2-mercaptoethanol (GIBCO ref. 31350010) and 100U/ml leukemia inhibitory factor (LIF, batch tested).

To differentiate ESCs into neurons, we performed monolayer neuronal differentiation (Aubert et al., 2002). ESCs were washed with PBS, dissociated using Accutase (StemPro ref. A1110501) and resuspended in N2B27 medium: 1:1 mix of Advanced DMEM/F-12 (GIBCO ref. 12634010) and Neurobasal (GIBCO ref. 21103049) supplemented with 1x L-Glutamine (GIBCO ref. 25030024), 1x MEM non-essential amino acids (GIBCO ref.11140035), 0.5x N-2 supplement (GIBCO ref. 17502048), 0.5x B-27 Supplement (GIBCO ref. 17504044) and 0.1mM 2-mercaptoethanol (GIBCO ref. 31350010). The appropriate number of cells (100,000 cells per well of a 6-well plate) was transferred into gelatin-coated plates containing N2B27 medium. The medium was changed every 2 days until analysis.

To assess self-renewal efficiency, ESCs were plated at clonal density (600 cells per well of a 6-well plate) in matrigel-coated (Corning ref. 354277) plates with N2B27 medium (see composition above) supplemented with "2i" inhibitors (Ying et al., 2008) (1μM PD0325901 (Axon ref. 1408) and 3μM CHIR99021 (Axon ref. 1386)) and 100U/ml LIF. Following 7 days of culture, cells were fixed and stained for alkaline phosphatase activity (AP) following manufacturer's instructions (Sigma-Aldrich ref. 86R-1KT). AP-positive colonies were imaged using a brightfield microscope (Nikon Ti2) and automatically counted using the ImageJ software.

Genetic manipulation of ESCs

To mutate endogenous *Sall4* genomic loci (*ZFC4mut*, *ZFC4Δ* and *ZFC1-2Δ*), E14Ju09 ESCs were modified by CRISPR/Cas9 (Ran et al., 2013). Guide RNAs were designed close to the desired mutation site (https://zlab.bio/guide-design-resources) and cloned into Cas9/gRNA co-expression plasmids (Addgene pX330, or derivative containing EGFP or a puromycin resistance cassette). Single-stranded repair DNA templates (ssDNAs) were ordered from Integrated DNA Technologies. ESCs ($4x10^5$ cells) were transfected with one (for point mutations) or two (for deletions) Cas9/gRNA plasmids and 10nmol of ssDNA template as appropriate. If a puromycin resistance cassette was used, cells were selected with puromycin and seeded at clonal density. If a fluorescent reporter was used, single cells were FACS-sorted and plated into wells of 96-well plates. ESC clones were expanded and their genomic DNA was extracted for genotyping by PCR (see primers) and Sanger sequencing.

To generate cell lines expressing a transgene of interest (Sall4 or EGFP cDNA) under a doxycycline-inducible promoter, *Sall4* knockout ESCs were modified using the PiggyBac (PB) transposon system. $1x10^6$ *Sall4* knockout ESCs were transfected with two PiggyBac vectors ("PB-$(TetO)_8$-Sall4-PGK-Hygromycin^R" or "PB-$(TetO)_8$-EGFP-PGK-Hygromycin^R" + "PB-Tet-On 3G-IRES-Zeocin^R"), together with a third plasmid expressing hyperactive PB transposase (Yusa et al., 2011) (pCMV-hyPBase). Approximately 48h post-transfection, ESCs were selected for 12 days with 200μg/ml hygromycin (doxycycline-inducible SALL4 or EGFP constructs) and 100μg/ml zeocin (Tet-On 3G transactivator construct). This experiment was repeated three times to obtain independent replicates for each cell line (SALL4 or EGFP). During selection, no doxycycline was added to the medium in order to prevent SALL4 or EGFP expression. To induce SALL4 or EGFP expression, cells were treated for 48h with 1μg/ml doxycycline (freshly prepared). For each replicate, SALL4 expression with and without doxycycline was controlled by RT-qPCR and immunofluorescence, as described below.

METHOD DETAILS

DNA pulldown and mass spectrometry

SILAC culture, preparation of heavy/light labeled nuclear protein extracts, DNA pulldowns and mass spectrometry were performed according to a previously published protocol (Spruijt et al., 2013b), with minor changes. Biotinylated bait (AT-run) and control (disrupted AT-run) DNA oligonucleotides (see Table S4) were purchased from Sigma-Aldrich and annealed as described. Poly(dI-dC) (Sigma-Aldrich ref. P4929) was used as competitor. Heavy- and light-labeled mouse ESC protein extracts were incubated with immobilized biotinylated DNA probes. After incubation and washes, beads from both pulldowns were combined after which bound proteins were digested with trypsin. Finally, the heavy/light ratio for the tryptic peptides was measured by LC-MS. Two replicate DNA pulldown/mass spectrometry experiments were performed with both bait/control pairs. The first experiment was done according to protocol using magnetic Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific ref. 65001) and in-gel digestion of samples after elution. In the second replicate experiment, agarose streptavidin beads (Thermo Fisher Scientific) were used and samples were digested on-beads prior to elution. Peptides were concentrated and desalted using StageTips (Rappsilber et al., 2003), before being analyzed on an EASY-nLC (Thermo Fisher Scientific) connected online to an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Peptides were measured during a 120min acetonitrile gradient using CID fragmentation of the top 15 precursor ions, with a dynamic exclusion duration of 30sec. Raw data was analyzed using MaxQuant (Cox and Mann, 2008) version 1.3.0.5. Using Perseus (Tyanova et al., 2016), the data was filtered for contaminants, reverse hits and the number of (unique) peptides. A scatterplot of the filtered data was generated using R. Detailed results from mass spectrometry analyses are available in Table S1.

Molecular Cell
Article

CellPress
OPEN ACCESS

DNA pulldowns for subsequent western blot analysis (see below) required scaling down of oligonucleotides, beads, Poly(dI-dC) competitor and total buffer volumes for use with 100μg or 200μg of nuclear protein extract. After binding of DNA oligonucleotides and washes with DNA binding buffer, beads were washed twice with protein binding buffer containing 0.5% BSA and blocked for 15min at room temperature. After incubation with nuclear protein extract, beads were washed five times in protein binding buffer and bound proteins were eluted by incubating beads in 50μl of NuPAGE LDS Sample Buffer (Thermo Fisher Scientific) for 15min at 70°C.

Immunoprecipitation

To prepare protein extracts for immunoprecipitation, ESCs were washed with PBS, trypsinised and collected in 15ml tubes. Following a centrifugation for 5min at 1,300rpm, the supernatant was removed and the cell pellet was resuspended in 1ml of lysis buffer (10mM NaCl, 1mM $MgCl_2$, 20mM HEPES pH7.5, 0.1% (v/v) Triton X-100) freshly supplemented with 1x protease inhibitor cocktail (Roche ref. 11873580001) and 0.5mM DTT. After a 20min incubation on ice with occasional shaking, nuclei were pelleted by centrifugation at 4°C for 10min at 1,500rpm. Supernatant was removed and nuclei were resuspended in 1ml of lysis buffer freshly supplemented with 1x protease inhibitor cocktail and 0.5mM DTT. The material was transferred into 1.5ml LoBind tubes (Eppendorf) and supplemented with 250U of Benzonase nuclease (Sigma-Aldrich). After a 5min incubation at room temperature, samples were supplemented with NaCl to obtain a final concentration of 150mM NaCl. Samples were incubated on a rotating wheel for 30min at 4°C. Tubes were centrifuged at 4°C for 30min at 13,300rpm and supernatants (nuclear protein extracts) were transferred into new 1.5ml LoBind tubes. 50μl of nuclear protein extract was boiled for 5min at 90°C in 2x Laemmli buffer (Sigma-Aldrich ref. S3401) as input material. Nuclear extracts were used directly for immunoprecipitation or stored at −80°C.

For immunoprecipitations, 5μg of anti-SALL4 antibody (Abcam ref. ab29112, RRID:AB_777810) was added to each nuclear protein extract (*Sall4* knockout protein extracts were used as negative control). Samples were incubated overnight at 4°C on a rotating wheel. 30μl of nProteinA Sepharose beads (GE Healthcare 4 Fast Flow), previously blocked with 0.5mg/ml BSA, were added to each nuclear extract and samples were incubated for 2h at 4°C on a rotating wheel. Samples were washed 5 times in lysis buffer freshly supplemented with 0.5mM DTT. Between each wash, samples were centrifuged at 4°C for 1min at 2,000rpm. After the final wash, beads were boiled for 5min at 90°C in 2x Laemmli buffer (Sigma-Aldrich ref. S3401) to elute the immunoprecipitated material.

Western blot

For western blot analysis, samples were loaded into 4%–15% Mini-PROTEAN TGX Precast gels (Bio-Rad), together with a fluorescent protein ladder (LI-COR ref. 928-60000). Proteins were separated by electrophoresis in SDS running buffer for ~45min at 200V. Subsequently, proteins were transferred on a nitrocellulose membrane at 4°C overnight at 23V. The membrane was blocked for 1h at room temperature with PBS supplemented with 10% non-fat skimmed milk and 0.1% Tween. The membrane was then incubated for 90min at room temperature with primary antibodies (see Table S4) diluted at the appropriate concentration in PBS supplemented with 5% non-fat skimmed milk and 0.1% Tween. The membrane was washed 4 times with PBS supplemented with 0.1% Tween, and incubated for 2h at room temperature with fluorescently labeled (LI-COR IRDye) or HRP-conjugated (GE Healthcare) secondary antibodies diluted in PBS supplemented with 5% non-fat skimmed milk and 0.1% Tween. The membrane was finally washed 4 times with PBS supplemented with 0.1% Tween. Proteins were visualized using the LI-COR Odyssey CLx imaging system (fluorescence) or detected on film by chemiluminescence (PerkinElmer ECL kit). Western blot signal was quantified using the LI-COR Image Studio software by measuring the fluorescence intensity of appropriate protein bands.

Immunofluorescence

For high resolution imaging, cells were plated on chambered coverslips (Ibidi ref. 80286). For lower magnification, cells were grown on standard tissue culture dishes. Cells were washed with PBS and fixed with 4% PFA for 10min at room temperature. After fixation, cells were washed with PBS and permeabilised for 10min at room temperature in PBS supplemented with 0.3% (v/v) Triton X-100. Samples were blocked for 1h30min at room temperature in blocking buffer: PBS supplemented with 0.1% (v/v) Triton X-100, 1% (w/v) BSA and 3% (v/v) serum of the same species as secondary antibodies were raised in (ordered from Sigma-Aldrich). Following blocking, samples were incubated overnight at 4°C with primary antibodies (see Table S4) diluted at the appropriate concentration in blocking buffer. After 4 washes in PBS supplemented with 0.1% (v/v) Triton X-100, samples were incubated for 2h at room temperature (in the dark) with fluorescently labeled secondary antibodies (Invitrogen Alexa Fluor Plus antibodies) diluted 1:1,000 in blocking buffer. Cells were washed 4 times with PBS supplemented with 0.1% (v/v) Triton X-100. DNA was stained with 4',6-diamidino-2-phenylindole (DAPI) for 5min at room temperature, and cells were submitted to a final wash with PBS. Samples were imaged by fluorescence microscopy (Nikon Ti2 or Zeiss LSM 880 with Airyscan). Images were analyzed and processed using the software Fiji.

RT-qPCR

Cells were directly lysed on the plate and total RNA was isolated using the RNeasy Plus Mini kit (QIAGEN ref. 74136), following manufacturer's instructions. The quantity and purity of RNA samples were determined using a micro-volume spectrophotometer (Nanodrop ND-1000). RNA was reverse-transcribed with SuperScript IV and random hexamers (Invitrogen ref. 18091050), following manufacturer's instructions. Triplicate qPCR reactions were set up in 384-well plates using the Takyon SYBR Mastermix (Eurogentec ref. UF-NSMT-B0701) and appropriate primer pairs (see Table S4). qPCR was performed and analyzed using the Roche LightCycler

CellPress OPEN ACCESS

Molecular Cell
Article

480 machine. For each primer pair, a standard curve was performed to assess amplification efficiency and melting curves were analyzed to verify the production of single DNA species.

HT-Selex

SELEX coupled with high-throughput sequencing (HT-SELEX) was performed as previously described (Jolma et al., 2010; Nitta et al., 2015), in three independent replicate experiments. All oligonucleotides were ordered from Integrated DNA Technologies (see Table S4). SELEX libraries were generated by PCR and consisted of 20bp random sequences flanked by primer binding sites for amplification. Double-stranded DNA libraries were purified using the MinElute PCR Purification Kit (QIAGEN ref. 28004) and controlled on a 10% polyacrylamide gel.

For SELEX experiments, recombinant SALL4 ZFC1 (residues 320-486), ZFC2 (residues 551-662) or ZFC4 (residues 859-1028) with an N-terminal hexahistidine tag were expressed from a pET-based vector in *E.coli* BL21 (*DE3*) cells. Proteins were purified using a 5 mL Histrap FF column, followed by separation by ion exchange (6 mL ResS column) and size exclusion chromatography (Superdex 200 16/600, all columns from GE Healthcare). SELEX libraries (1.5μg for the first cycle, 200ng for subsequent cycles) were mixed with 1μg of hexahistidine-tagged SALL4 ZFC in 100μl of SELEX buffer (50mM NaCl, 1mM $MgCl_2$, 0.5mM EDTA, 10mM Tris-HCl pH7.5, 4% glycerol) freshly supplemented with 5μg/ml Poly(dI-dC) and 0.5mM DTT. A negative control experiment (without addition of proteins) was also performed to control for technical bias during the SELEX protocol. Following a 10min incubation at room temperature on a rotating wheel, 50μl of Ni Sepharose 6 Fast Flow beads (GE Healthcare), previously equilibrated in SELEX buffer, were added to each sample and incubated for an additional 20min at room temperature on a rotating wheel. To remove non-specifically bound oligonucleotides, beads were washed 5 times with 1ml of SELEX buffer, freshly supplemented with 0.5mM DTT. Between each wash, samples were incubated for 5min at room temperature on a rotating wheel and centrifuged for 1min at 2,000 rpm. After the final wash, beads were resuspended in 100μl H_2O and used directly for PCR using the high-fidelity Phusion DNA polymerase (NEB ref. M0530S). The minimum number of PCR cycles required to amplify each library was determined by running samples amplified with increasing PCR cycle numbers on 10% polyacrylamide gels. Amplified libraries were purified using the MinElute PCR Purification Kit and used for subsequent rounds of SELEX. To generate libraries for high-throughput sequencing at the desired number of SELEX cycles, libraries were amplified using primers containing Illumina adapters and unique barcodes. Double-stranded DNA libraries were purified using the QIAquick PCR Purification Kit (QIAGEN ref. 28104). Contaminating primers were eliminated by size exclusion using KAPA Pure beads (Roche ref. 07983271001) with a 3x beads-to-sample ratio. SELEX libraries with unique barcodes were pooled in equimolar amounts and sequenced using the Illumina MiSeq platform (EMBL GeneCore facility, Germany).

ChIP-seq

ChIP was performed as previously described (Stock et al., 2007), in two or three independent replicate experiments for each sample. For each ChIP, $25x10^6$ ESCs were plated into 15cm dishes the day before the experiment. Cells were crosslinked at 37°C for 10min with 1% formaldehyde. Following quenching for 5min at room temperature with 125mM glycine, cells were washed 3 times with ice-cold PBS. Swelling buffer (10ml of 10mM KCl, 1.5mM $MgCl_2$, 25mM HEPES pH7.9, 0.1% NP-40) freshly supplemented with 1x protease inhibitor cocktail (Roche ref. 11873580001) was added into each plate, followed by a 10min incubation at 4°C. Nuclei were collected by scraping and transferred into 15ml tubes. Samples were centrifuged at 4°C for 5min at 3,000rpm and the supernatant was removed. Crosslinked nuclei were quickly frozen on dry ice and stored at −80°C. Crosslinked nuclei were thawed on ice, resuspended in 2ml of sonication buffer (140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, 50mM HEPES pH7.9) freshly supplemented with 1x protease inhibitor cocktail, and transferred into 1.5ml TPX tubes (Diagenode). Chromatin was sonicated by performing 20x sonication cycles (30sec on/ 30sec off) using the Bioruptor Twin instrument (Diagenode) with a 4°C water bath. Samples were centrifuged at 4°C for 30min at 13,000rpm to remove insoluble material. Supernatants (soluble chromatin fraction) were collected and transferred into 1.5ml LoBind tubes (Eppendorf). To evaluate the amount of chromatin in each sample, a 2μl aliquot was alkaline-lysed with 0.1M NaOH and measured using a micro-volume spectrophotometer (Nanodrop ND-1000).

For each immunoprecipitation, 700μg of chromatin was mixed with 5μg of anti-SALL4 antibody (Santa Cruz ref. sc-101147, RRID:AB_1129262 or Abcam ref. ab29112, RRID:AB_777810) in a total volume of 1ml of sonication buffer supplemented with 1x protease inhibitor cocktail. *Sall4* knockout ESCs chromatin samples were used as a negative control. Samples were incubated overnight at 4°C on a rotating wheel. 50μl of either Protein A (ChIP with Abcam ref. ab29112) or Protein G (ChIP with Santa Cruz ref. sc-101147) magnetic beads (Invitrogen Dynabeads), previously equilibrated in sonication buffer, was added into each sample. Following a 3h incubation at 4°C on a rotating wheel, beads were extensively washed with 1ml of each of the following buffer: 1x with sonication buffer, 1x with wash buffer A (500mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, 50mM HEPES pH7.9), 1x with wash buffer B (250mM LiCl, 1mM EDTA, 0.5% NP-40, 0.5% Na-deoxycholate, 20mM Tris pH8.0), 2x with TE buffer (Sigma-Aldrich ref. 93283). Between each wash, beads were incubated for 5min at room temperature on a rotating wheel. Finally, DNA was eluted by resuspending beads in 250μl of elution buffer (50mM Tris pH7.5, 1mM EDTA) freshly supplemented with 1% SDS, and by incubating samples at 65°C for 5 min. Samples were further incubated for 15min at room temperature on a rotating wheel and the supernatant (eluted chromatin) was collected into a new 1.5ml LoBind tube. The elution was repeated a second time to obtain 500μl of immunoprecipitated chromatin.

To extract DNA from immunoprecipitated chromatin or from the input material (50μl of soluble chromatin), crosslinking was reversed by incubating samples overnight at 65°C in a total volume of 500μl with 160mM NaCl and 20 μg/ml RNase A. Then,

Molecular Cell
Article

CellPress
OPEN ACCESS

5mM EDTA and 200 μg/ml Proteinase K were added to the samples, followed by a 2h incubation at 45°C. Finally, DNA was purified by phenol-chloroform extraction (Invitrogen ref. 15593031) followed by ethanol precipitation with 2x volumes of 100% ethanol, 0.1x volume of 3M sodium acetate, and 40μg of glycogen (Invitrogen ref. 10814010). Samples were incubated at −80°C for at least 1h and centrifuged at 4°C for 30min at 13,000rpm. The supernatant was removed and DNA pellets were washed with 70% EtOH. Following a final spin at 4°C for 15min at 13,000rpm, DNA pellets were air-dried and resuspended in 30-100μl TE buffer (Sigma-Aldrich ref. 93283) or H_2O. DNA concentration was quantified using the Qubit dsDNA HS Assay Kit (Invitrogen ref. Q32854).

ChIP-seq libraries were prepared using the KAPA Hyperprep Kit (Roche ref. 07962347001) together with KAPA dual-indexed adapters (Roche ref. 08278555702), following manufacturer's instruction. ChIP-seq libraries were quantified using the Qubit dsDNA HS Assay Kit (Invitrogen ref. Q32854) and fragment size was evaluated using the Agilent 2100 Bioanalyzer (Agilent High Sensitivity DNA Kit). ChIP-seq libraries with unique barcodes were pooled in equimolar amounts and sequenced using the Illumina HiSeq 4000 and NextSeq 500 platforms (EMBL GeneCore facility, Germany).

ATAC-seq

ATAC-seq was performed as previously described (Cholewa-Waclaw et al., 2019). ESC nuclei from three independent *WT* ESC replicates were isolated using hypotonic buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM $MgCl_2$, 0.1% Igepal CA-630). 50,000 nuclei were resuspended in 50μl of 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°C. DNA was purified and PCR amplified with the NEBNext High Fidelity reaction mix (NEB) to generate DNA libraries. Libraries were sequenced using the Illumina HiSeq 2500 platform with 75bp paired-end sequencing.

RNA-seq

For RNA-seq in ESCs, all cell lines were seeded at the same density in 6-well plates, in three or four independent replicate experiments for each sample. Following two days of culture, total RNA was extracted using the AllPrep DNA/RNA kit (QIAGEN) or the RNeasy Plus Mini kit (QIAGEN), following the manufacturer's instructions and contaminating genomic DNA was removed by DNase I treatment. Before library preparation, equal amounts of either RNA sequins (Garvan Institute of Medical Research, Australia) or ERCC (Invitrogen) spike-in mix were added to each sample. Ribosomal RNA-depleted RNA-seq libraries were prepared using either the ScriptSeq Complete Gold Kit (Illumina) or the KAPA RNA Hyperprep Kit (Roche ref. 08098131702) together with indexed adapters, following the manufacturer's instructions. RNA-seq libraries with unique barcodes were pooled in equimolar amounts and sequenced using the Illumina HiSeq 2500 (Wellcome Sanger Institute, UK), HiSeq X (Novogene Europe, UK) or NextSeq 500 (EMBL GeneCore facility, Germany) platforms.

For the RNA-seq time course experiment, cells were submitted to neuronal differentiation as previously described (see cell culture section), in two independent replicate experiments for each sample. At the appropriate time point, cells were directly lysed on the plate and total RNA was extracted using the RNeasy Plus Mini kit (QIAGEN), following the manufacturer's instructions and contaminating genomic DNA was removed by DNase I treatment. Equal amounts of RNA sequins spike-in mix (Garvan Institute of Medical Research, Australia) were added to each sample and RNA-seq libraries were prepared by polyA-enrichment using the NEBNext Ultra II Library Prep Kit (NEB ref. E7645) together with indexed adapters. RNA-seq libraries with unique barcodes were pooled in equimolar amounts and sequenced using the Illumina NovaSeq platform (Novogene Europe, UK).

QUANTIFICATION AND STATISTICAL ANALYSIS

HT-SELEX analysis

All possible k-mers (width = 5) were searched individually in all SELEX libraries at different cycles. The fraction of reads containing the k-mer was considered as its abundance. Subsequently, top 3 abundant k-mers from ZFC4 SELEX library at cycle 6 were searched allowing one mismatch and a position frequency matrix (PFM) was generated for each. The PFM was used to visualize the motifs using Logolas (Dey et al., 2018).

RNA-seq analysis

Alignment-free quantification from RNA-seq data was performed using sailfish v0.9.2 (Patro et al., 2014). Annotation data was downloaded from Gencode and a transcriptome index was generated for assembly release M23. Differential gene expression analysis was performed using DESeq2 v1.28.0 (Love et al., 2014) and genes with adjusted p value < 0.05 were considered for further analyses. Genome wide base composition was calculated for 1 kilobase (kb) windows of the mouse genome using bedtools nuc (Quinlan and Hall, 2010) and the AT content BigWig track was generated. Base composition for multiple gene loci was calculated using deepTools computeMatrix (Ramírez et al., 2014). Gene ontology analysis for genes deregulated in *ZFC4mut/Δ* ESCs was performed using clusterProfiler Bioconductor package (Yu et al., 2012) and simplified GO terms from enrichGO function were used to identify enriched GO Terms with q-value < 0.01 as significance threshold (see Table S3). Bootstrapped two-sided t tests were used to associate statistical significance for comparisons of $\log_2$ fold changes in different *Sall4* mutants compared to wild-type (*WT*) for neuronal differentiation genes. Command line arguments and source code for analysis is detailed in Methods S1.

CellPress
OPEN ACCESS

Molecular Cell
Article

ChIP-seq analysis

Sequencing reads were trimmed using Trimmomatic v0.33 (Bolger et al., 2014) and aligned to mm10 assembly using bwa-mem v0.7.17 (Li, 2013). PCR duplicate sequencing reads were removed using MarkDuplicates from Picard toolkit (http://broadinstitute.github.io/picard/). GC-bias was estimated for input chromatin samples using computeGCBias (Benjamini and Speed, 2012) from deepTools (Ramírez et al., 2014). Subsequently, both ChIP and input chromatin samples were corrected using the input chromatin estimated bias using correctGCBias. Peak calling on the GC-bias corrected BAM files was performed using MACS v2.1.2 (Zhang et al., 2008). BigWig tracks for ChIP over input chromatin were calculated using bamCompare. For meta-analyses of peak regions, ChIP signal scores per genome regions was calculated using computeMatrix. Motif discovery and motif enrichment analysis was performed using MEME-ChIP v5.1.0 (Machanick and Bailey, 2011) for ChIP-seq peaks with background sequences randomly sampled from accessible chromatin regions. Command line arguments and source code for analysis is detailed in Methods S1.

ATAC-seq analysis

Sequencing reads were trimmed using Trimmomatic v0.33 (Bolger et al., 2014) and aligned to mm10 assembly using bwa-mem v0.7.17 (Li, 2013). PCR duplicate sequencing reads were removed using MarkDuplicates from Picard toolkit (http://broadinstitute.github.io/picard/). Broad peak calling on the de-duplicated BAM files was performed using MACS v2.1.2 (Zhang et al., 2008).

Quantification of AT effect

Ordinary least-squares (OLS) linear regression was fitted by selecting RNA-seq $\log_2$ fold change as an endogenous variable and average AT content across the gene locus as an exogenous variable. For every model fit, the p value associated with the F-statistic and quantified AT effect with a confidence interval of 99% was used for further analysis. R^2 values were estimated from a linear regression fit when $\log_2$ fold change is regressed against AT content across gene locus. p values obtained from all model fits were adjusted using the Benjamini/Hochberg multiple testing comparison and models with a false discovery rate (FDR) < 0.01 were deemed significant. Detailed results from statistical analyses are available in Table S2. Command line arguments and source code for analysis is detailed in Methods S1.

5.3 Conclusion

Our findings demonstrate that the extent of SALL4-mediated transcriptional repression is linearly correlated with gene body AT content, reinforcing the concept that sequence composition serves as a regulatory determinant. Through quantitative modelling, we show that genes with higher AT composition are more susceptible to SALL4-mediated repression, consistent with a model where SALL4 acts as a selective reader of DNA sequence architecture to modulate gene expression programs.

Chapter 6

Interpreting epigenetic regulation of gene expression

As outlined in Section 3.3, the regulation of transcription involves intricate interactions between chromatin-associated proteins, yet identifying their precise contributions remains a challenge. We test the hypothesis that explainable artificial intelligence (XAI) methods, specifically SHAP, can offer insights in transcriptional regulation. The experiments performed to evaluate this approach, along with their results, are presented as they appear in my preprint (Chhatbar et al. 2025) undergoing revision at PLOS Genetics.

6.1 Introduction

The regulation of transcription is a complex interplay between chromatin-associated proteins, yet deciphering their precise mechanistic contributions remains a challenge. Advances in machine learning and XAI offer powerful tools to uncover these relationships by integrating computational modelling with functional genomics data. This study employs deep learning models to predict Pol–II occupancy based on chromatin-associated protein profiles and assesses the effectiveness of SHAP in inferring regulatory mechanisms. By leveraging degron-based perturbation datasets, we explore the potential of SHAP to highlight regulatory relationships and suggest possible mechanisms underlying gene expression changes.

6.2 Modeling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies

Modelling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies

Kashyap Chhatbar[1,2], Adrian Bird[2], Guido Sanguinetti[3]

1 School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
2 Institute of Cell Biology, University of Edinburgh, Edinburgh, United Kingdom
3 International School for Advanced Studies (SISSA), Trieste, Italy

* kashyap.c@ed.ac.uk

Abstract

Transcriptional regulation involves complex interactions with chromatin-associated proteins, but disentangling these mechanistically remains challenging. Here, we generate deep learning models to predict RNA Pol-II occupancy from chromatin-associated protein profiles in unperturbed conditions. We evaluate the suitability of Shapley Additive Explanations (SHAP), a widely used explainable AI (XAI) approach, to infer functional relevance and analyse regulatory mechanisms across diverse datasets. We aim to validate these insights using data from degron-based perturbation experiments. Remarkably, genes ranked by SHAP importance predict direct targets of perturbation even from unperturbed data, enabling inference without costly experimental interventions. Our analysis reveals that SHAP not only predicts differential gene expression but also captures the magnitude of transcriptional changes. We validate the cooperative roles of SET1A and ZC3H4 at promoters and uncover novel regulatory contributions of ZC3H4 at gene bodies in influencing transcription. Cross-dataset validation uncovers unexpected connections between ZC3H4, a component of the Restrictor complex, and INTS11, part of the Integrator complex, suggesting crosstalk mediated by H3K4me3 and the SET1/COMPASS complex in transcriptional regulation. These findings highlight the power of integrating predictive modelling and experimental validation to unravel complex context-dependent regulatory networks and generate novel biological hypotheses.

Author summary

Genes are turned on or off through complex processes involving many proteins that interact with DNA wrapped histones and modify their structure. These changes, known as epigenetic modifications, help control how genes are expressed without altering the DNA sequence itself. In this study, we wanted to understand how different proteins influence gene activity in mouse stem cells by looking at their positions along the genome, particularly whether they act near the gene's start site (promoter) or within the gene body. To do this, we used machine learning models and a method called SHAP, which helps explain the model's decisions. By comparing our predictions to data from experiments where specific proteins were removed, we found that some proteins have context-specific effects, acting not only at the promoter but also along the whole gene body. Our approach highlighted both well-known and unexpected regulators of transcription and revealed that gene body signals, which are often overlooked, can play key roles. These findings show how explainable AI can help uncover new insights into how epigenetic features shape gene regulation, and offer a powerful way to generate testable hypotheses from complex genomic data.

Introduction

Understanding how cells regulate gene expression during physiological or disease processes remains one of the most important open problems in biology, with potentially immense implications for biomedicine and biotechnology. While the fundamental process of regulation by transcription factor binding has been extensively studied [1], the complexity of sequence-dependent signal integration in a crowded chromatin environment remains largely unknown. Through

efforts to tackle this outstanding problem, the last decade has witnessed the development of several large-scale international efforts to profile multiple epigenomic marks across a variety of cellular contexts [2–4]. The large amounts of data made publicly available by these projects raised the prospect of a complementary, data-driven approach to understand general mechanisms by which gene expression is regulated. Numerous studies have used statistical models to predict gene expression or other molecular features from genomic sequence and epigenomic marks [5–11]. More recently, advances in artificial intelligence (AI) using a variety of neural architectures have been brought to bear, allowing increasingly accurate prediction and imputation of gene expression as well as other molecular features [12–15]. A critical limitation of this data-driven approach is its entirely correlative nature. Because the vast majority of data sets are derived from cells at steady-state, every prediction will demand independent validation via explicit perturbation experiments, a requirement that has rarely been met.

Since complex models generally trade off biological interpretability against statistical accuracy, so-called explainable AI (XAI) techniques are adopted to provide human interpretable explanations for specific model predictions and to hypothesize potential underpinning biological mechanisms [16–18]. In this paper, we set out to benchmark the efficacy of XAI approaches in identifying regulatory processes by taking advantage of the recent availability of dynamical measurements provided after rapid perturbations via auxin-inducible degron (AID) and degradation tag (dTAG) systems [19]. We set out to predict RNA Polymerase II occupancy from several chromatin associated proteins and epigenomic marks, quantifying the influence of each protein or mark on each gene using SHAP (Shapley Additive Explanations) [17], one of the most widely used XAI approaches. We then validate the efficacy of the learnt explanations by systematic comparisons of direct target and random gene sets following degradation of one or more factors. Our findings demonstrate the utility of combining predictive modelling with SHAP analysis to uncover complex biological interactions and infer functional mechanisms in transcriptional regulation. By leveraging this integrative approach, we dissected the roles of chromatin-associated proteins, uncovered unexpected functional relationships, and validated the robustness of SHAP across datasets.

Results

Data sets used for modelling

To ensure a robust and comprehensive analysis of transcriptional regulation, we utilised datasets from three independent studies spanning different experimental designs and chromatin-associated protein targets: Dobrinić et al. (2021) [20], Hughes et al. (2023) [21], and Wang et al. (2023) [22] (Table 1). Dobrinić and colleagues [20] investigated the roles of PRC1 [23,24] and PRC2 [25,26] complexes in Polycomb-mediated transcriptional repression, using ChIP-seq data for RING1B, SUZ12, and associated histone modifications such as H3K27me3 and H2AK119ub1 to evaluate their contributions. Additionally, this study included degron-based perturbation experiments for RING1B and SUZ12, enabling a direct assessment of acute transcriptional changes using nuclear RNA-seq assays. Hughes and colleagues [21] focused on transcriptional regulation mediated by SET1A [27] and ZC3H4 [28], two chromatin-associated proteins with well-characterised roles in activating and restricting transcription, respectively. This study leveraged ChIP-seq data for SET1A, ZC3H4, and H3K4me3, alongside degron-based perturbation experiments, to uncover the interplay between these factors to regulate transcription. Wang and colleagues [22] explored the role of H3K4me3 in transcription initiation and RNA Pol-II pause-release, utilizing ChIP-seq data for the SET1 complex components DPY30 and RBBP5 [27], Integrator complex member INTS11 [29,30], and factors involved in transcription elongation and Pol-II pause-release such as CDK9 [31], BRD4 [32], HEXIM1, NELFA [33], and AFF4 [34]. Transcriptional dynamics after acute perturbations were assayed using TT-chem-seq [21,22]. Together, these datasets provided a rich resource for examining the relationships between chromatin modifications and transcriptional regulation.

Modelling of protein occupancy predicts transcription output

In order to understand the relationships between protein occupancy profiles and transcriptional regulation, we built deep neural network (DNN) models, specifically multilayer perceptrons (MLPs), gradient-boosted trees (XGBoost) and linear regression to predict RNA Pol-II occupancy from ChIP-seq data (Fig 1A). MLPs are prioritised in this study for

their competitive performance and efficiency in regression tasks [35, 36], along with their scalability and potential for improved interpretability using XAI approaches. Since epigenomic data is not necessarily statistically independent, we also utilised gradient-boosted trees. To enhance the resolution of our predictions, normalised ChIP signals were quantified separately from promoter regions and gene body regions, enabling us to investigate the distinct contributions of chromatin proteins and marks to transcription regulation across these regions. By integrating these separate measurements, we aimed to identify specific factors or chromatin features that predominantly influence transcription at promoters or gene bodies. For all data sets, the occupancy of RNA Pol-II was predicted with high precision (R^2 ranging between 0.85–0.95) (Fig 1B). Model performance was assessed with shuffled 5-fold KFold cross-validation.

We hypothesise that the model, trained on data assayed from unperturbed conditions, captures biologically meaningful relationships between protein occupancy and transcriptional output. To interpret these relationships, we use SHAP (SHapley Additive exPlanations) [17, 37], an XAI approach that quantifies the contribution of each input feature to the model's predictions while potentially accounting for complex, non-linear interactions (Fig 1A). After we modelled the data using MLP and XGBoost regression, we utilise KernelSHAP, DeepSHAP and TreeSHAP algorithms from the SHAP library. Through comparing SHAP values between direct transcriptional targets (genes that exhibit significant transcriptional changes following rapid protein degradation) and random genes, we aim to assess the relative importance of specific protein occupancy features at promoter and gene body context in driving transcriptional outcomes. For robust comparison, we categorised the genes derived from data generated from nascent (TT-seq) or steady state RNA-seq experiments after acute degradation of the protein of interest. The categorisation was based on the significance threshold using p-adjusted values attributed to the confidence in estimating the fold change. We calculated the mean absolute SHAP values for each group and the error is estimated through random subsampling. Since the model is trained on data assayed before any perturbations, differences in SHAP values between direct targets (p-adjusted value < 0.05) and random genes (p-adjusted value > 0.05) may indicate SHAP's potential to identify functional relevance. After successful training, SHAP values are computed for gene loci present only in the test dataset to prevent data leakage. Because direct targets constitute 6–28% of genes in our degron datasets, every test split was built to preserve that proportion. For each split we drew 50 balanced subsamples, each containing 50 direct-target genes (p-adjusted value < 0.05) and 50 randomly selected non-targets (p-adjusted value > 0.05). Absolute SHAP values were averaged across these repeats to minimise bias from class imbalance.

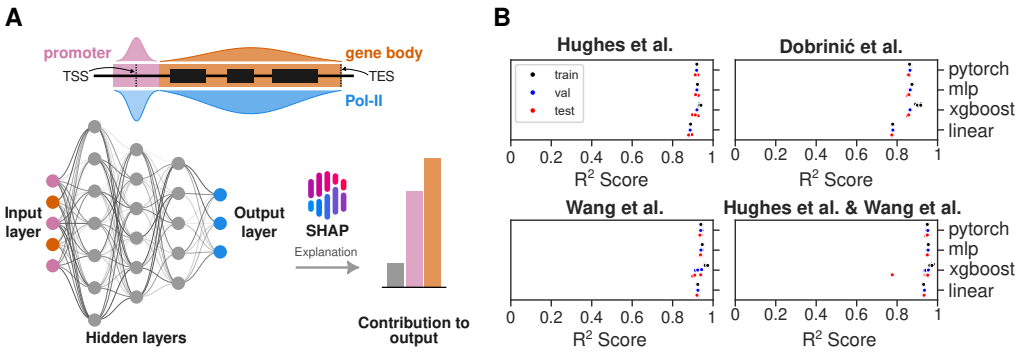


Fig 1. Overview of predictive modelling for RNA Pol-II occupancy. (A) Schematic of the multilayer perceptron (MLP) model architecture used to predict RNA Pol-II occupancy based on chromatin-associated protein profiles. Input features include protein occupancy data from ChIP-seq experiments, and outputs represent predicted RNA Pol-II occupancy. (B) Model performance for predicting RNA Pol-II occupancy based on chromatin-associated protein profiles across datasets. Coefficients of determination (R^2) are shown for training, validation and test datasets for 5 splits.

SHAP value is an estimate of the original Shapley value [37] which is a unique solution that fairly distributes the difference between the model prediction and a background expectation across all features. Because these estimations possess desirable properties of local accuracy, missingness and consistency, summing all values reconstructs the predicted RNA Pol-II occupancy for each gene. These surprising and unique attributes allows direct comparison of feature contributions on a gene-by-gene or gene-set-by-gene-set basis. To benchmark SHAP's effectiveness against traditional approaches, we also conducted separate linear regression analyses for target and random genes. To

establish a challenging benchmark, a linear regression baseline was employed, deliberately leveraging outcome-related information that would be inaccessible at inference time for the SHAP-enabled models. Linear regression coefficients can provide only global insights, revealing general trends across the entire set of target or random genes. In contrast, SHAP analysis calculates feature importance scores for each gene individually, offering unprecedented resolution into how different regulatory features influence transcription in a gene-specific manner. We anticipated that this gene-by-gene analysis would reveal complex, context-dependent interactions between features that would be masked by global linear trends. The comparison of absolute mean values for different gene sets whether they are direct targets or not allows estimating whether a particular mark or multiple marks contribute significantly in predicting RNA Pol-II occupancy. This approach was designed to test whether SHAP's ability to capture gene-specific regulatory patterns could provide deeper mechanistic insights than traditional analyses.

Strengths and limitations of SHAP analysis in resolving co-occupying transcriptional regulators

To evaluate SHAP's ability to accurately attribute importance to individual proteins in complex regulatory systems, we generated a highly accurate model (Fig 1B) to predict RNA Pol-II occupancy using PRC1, PRC2 and H3K4me3 ChIP-seq data as they were sequenced together in the Dobrinić et al. study [20] (Fig 2A). PRC1, through its core subunit RING1B [23,24], functions as the primary transcriptional repressor [38], while PRC2 reinforces repression by depositing the H3K27me3 histone mark [39]. The extensive co-occupancy of PRC1 and PRC2 at target loci [40] provided an ideal system to test SHAP's performance in disentangling overlapping regulatory inputs. As expected, RING1B promoter occupancy emerged as significantly more important at direct target gene loci (S1 FigA) when SHAP values were calculated using DeepSHAP algorithm. When we extended the analysis using KernelSHAP and TreeSHAP across all time points (2h, 4h, 8h, and 24h; S1 FigB) following RING1B depletion, we consistently identified RING1B as more important at direct targets. This robust finding aligns with extensive research that has established PRC1's dominant role in early transcriptional repression. This consistency represents a significant achievement, as it demonstrates SHAP's ability to capture biologically meaningful features, even in a highly co-linear regulatory environment. Across all time points, SHAP attributed greater importance to SUZ12 and its catalytic histone mark H3K27me3 at direct target genes of RING1B than at random genes, yet both still ranked below RING1B itself (S1 FigB). In all our SHAP explainers except TreeSHAP at 2h and 4h, RING1B was ranked higher than H3K27me3 (S1 FigB). It is important to note that the model was trained on unperturbed data, and thus it cannot reasonably infer or capture the temporal ordering of regulatory effects. The dynamics of feedback mechanism between PRC1 and PRC2 [25,41–43] makes it plausible for H3K27me3, which is deposited by SUZ12 to show higher importance at these time points.

Interestingly, SHAP also highlighted the importance of H3K4me3 at all gene loci regardless of perturbation following RING1B depletion (S1 FigA, B), raising the possibility that the SET1 complex, which deposits this histone mark [27] is recruited to polycomb-repressed genes. This observation aligns with the established roles of CFP1, a SET1 complex component that binds CpG islands (CGIs) [44], and KDM2B, a PRC1 complex recruiter which also binds CGIs [45]. These findings provide further evidence of a dynamic interplay between the SET1 and PRC complexes, wherein CFP1 and KDM2B continuously sample CGIs but ultimately yield to the reinforcing feedback mechanisms of PRC1 and PRC2 [46], maintaining transcriptionally repressed loci. These findings also raise the possibility that these loci represent "bivalent" promoters, where H3K4me3 and H3K27me3 coexist [47]. This dual marking is thought to safeguard proper and robust differentiation by maintaining genes in a poised state, balancing activation and repression until lineage-specific cues drive transcriptional commitment [47,48], further emphasizing the dynamic regulatory interplay revealed by our analysis. Another possibility is H3K4me3's well documented association with RNA Pol-II. We reasoned that this model is relying heavily on H3K4me3 to predict RNA Pol-II occupancy. Therefore, we trained new models without incorporating H3K4me3 occupancy. Surprisingly, these new models retained similar levels of predictive accuracy with test R^2 scores dropping from 0.85 (Fig 1B) to 0.8 (S2 FigA) indicating that the modelling strategy does not rely heavily on the correlative relationship between H3K4me3 and RNA Pol-II. When these updated models were interpreted using multiple interpretability algorithms, absolute SHAP values for RING1B promoter occupancy were consistently higher for target gene loci across all time points (Fig 2B, C), following RING1B depletion. Surprisingly, regression coefficients obtained from linear models failed to highlight any significant difference between direct targets and random genes across all time points (S2 FigB, C, D and E). These results demonstrate the utility of SHAP in consistently identifying biologically validated regulators like RING1B as key players, even in complex and co-linear

regulatory systems. We argue that our analysis of chromatin-associated proteins, where SHAP values are significantly higher for target genes compared to random genes, is analogous to assessing the functional impact of these proteins.

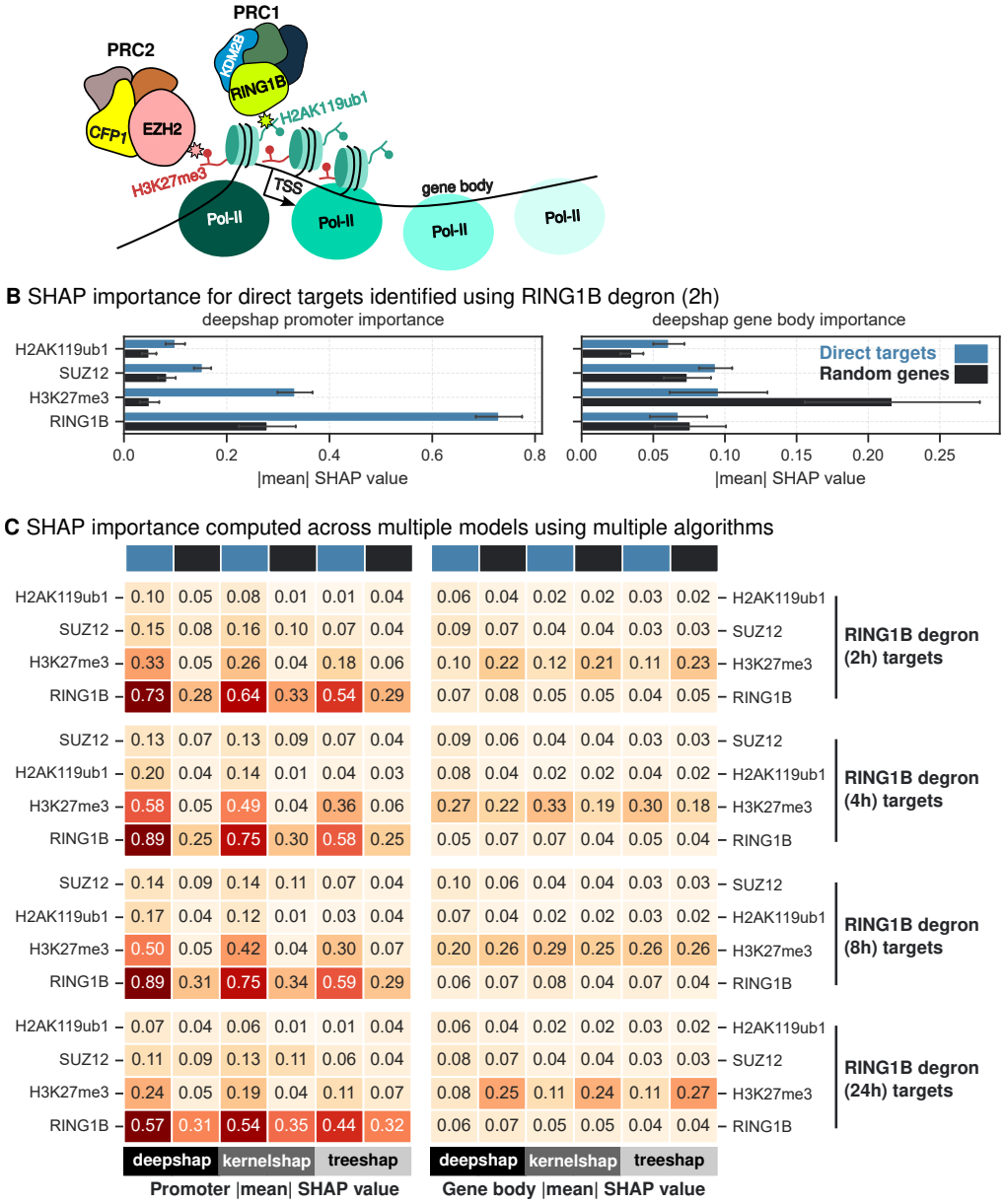


Fig 2. SHAP values for co-occupying PRC1 (RING1B) and PRC2 (SUZ12) components and their catalytic marks. (A) Chromatin-associated proteins and histone marks relevant for this model. (B) Comparison of absolute mean DeepSHAP values computed for direct targets (RING1B degron 2h) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. (C) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) across all time points following RING1B degradation (2h, 4h, 8h and 24h).

Interpretable modelling analysis aids in inferring the combined role of SET1A and ZC3H4 in transcriptional regulation

A recent study [21] revealed the combined role of SET1A and ZC3H4 in transcriptional regulation, highlighting their interplay. Building on these datasets, we aimed to determine whether XAI approaches could infer relationships between these proteins directly from predictive models trained on unperturbed data. Understanding such cooperative roles demands methodologies that go beyond traditional bioinformatic techniques, which often rely on summary statistics and correlation analyses to infer biological mechanisms. For instance, SET1A, a histone methyltransferase responsible for H3K4 trimethylation [27], co-occupies promoters with H3K4me3. SET1A, in complex with WDR82, functions as a transcriptional activator, while ZC3H4, also interacting with WDR82, is part of the restrictor complex that regulates non-coding transcription [28, 49].

To shed light on their roles in transcriptional regulation, we utilised our highly accurate model for Hughes et al. [21] data (Fig 1B) which predicts RNA Pol-II occupancy using the occupancy profiles of SET1A, ZC3H4, and H3K4me3 (Fig 3A). We performed SHAP analysis on gene sets following the acute degradation of SET1A, ZC3H4, and both proteins simultaneously. SHAP values were consistently higher for SET1A, ZC3H4 and H3K4me3 promoter occupancy in direct targets compared to random genes (Fig 3B), indicating that the model captures biologically relevant effects of these proteins on transcriptional regulation. Additionally, SHAP analysis highlighted a higher importance of ZC3H4 gene body occupancy in target genes, whereas SET1A showed no significant increase in gene body importance (Fig 3B). As a control comparison, we trained separate linear regression models on direct targets and random genes to evaluate whether the relationships between input features and transcriptional regulation differ across conditions. The coefficients derived from these models showed no significant differences (S3 FigC), indicating that the predictive relationships remain consistent between direct targets and random gene sets. These results demonstrate that XAI approaches, such as SHAP, provide valuable insights into the cooperative roles of SET1A and ZC3H4 at promoters compared to linear regression coefficients. When we extended the analysis using multiple SHAP algorithms and direct targets from ZC3H4 degron and SET1A+ZC3H4 degron, we consistently found SET1A, ZC3H4 and H3K4me3 important at promoter context and ZC3H4 at gene body context (Fig 3C, S3 FigA, B). The promoter context finding aligns with established experimental evidence [21], validating the robustness of our approach. Surprisingly, the distinct regulatory role of ZC3H4 at gene body represents a potentially novel phenomenon, extending beyond published conclusions and uncovering novel aspects of ZC3H4's contribution to transcriptional regulation.

SHAP analysis validates the importance of INTS11 at target genes

INTS11, a core component of the Integrator complex [29], is recruited to active promoters through H3K4me3 [22]. This interaction highlights the crosstalk between the SET1/COMPASS complex and Integrator complex in regulating transcriptional activity (Fig 4A). To further investigate these relationships, we utilised our model, employing SHAP to quantify the contributions of DPY30, RBBP5 and INTS11 along with CDK9, HEXIM1, BRD4, NELFA, and AFF4, each playing crucial roles in transcription elongation and Pol-II pause-release [31, 32, 34, 50]. When we utilised our model trained on data from Wang et al. [22] (Fig 1B), our DeepSHAP analysis showed that INTS11 and H3K4me3 promoter occupancy exhibited high SHAP values at target gene loci, regardless of whether INTS11 (Fig 4B), DPY30 (S4 FigA), or RBBP5 (S4 FigB) was degraded. This finding underscores INTS11's central role, likely facilitated through its interaction with H3K4me3 at active promoters. Notably, H3K4me3 promoter occupancy was also identified as significant contributors to RNA Pol-II occupancy, particularly at INTS11 target gene loci (Fig 4B). Similar to our previous observations, coefficients derived from linear regression models showed little differences (S4 FigC) between models trained separately for direct targets (INTS11 degron) and random genes whereas SHAP values computed from multiple algorithms clearly separated target genes from random genes (Fig 4B, C). Despite DPY30 and RBBP5 being part of the SET1/COMPASS complex depositing H3K4me3, SHAP analysis revealed a lack of significant importance for DPY30 and RBBP5 (Fig 4B, S4 FigA, B), an observation that stands in contrast to expectations. We also noticed that INTS11 SHAP values for direct targets of DPY30 degron (S4 FigA) and RBBP5 degron (S4 FigB) showed only marginal increase compared to random genes. Nonetheless, our analysis highlights the importance of H3K4me3 promoter occupancy in the regulation of INTS11 target genes (Fig 4B, C), supporting its in regulating RNA Pol-II pausing [22]. Our results independently validate the central role of INTS11 in regulating transcription at target loci, emphasizing its critical function in transcriptional regulation through H3K4me3 recruitment.

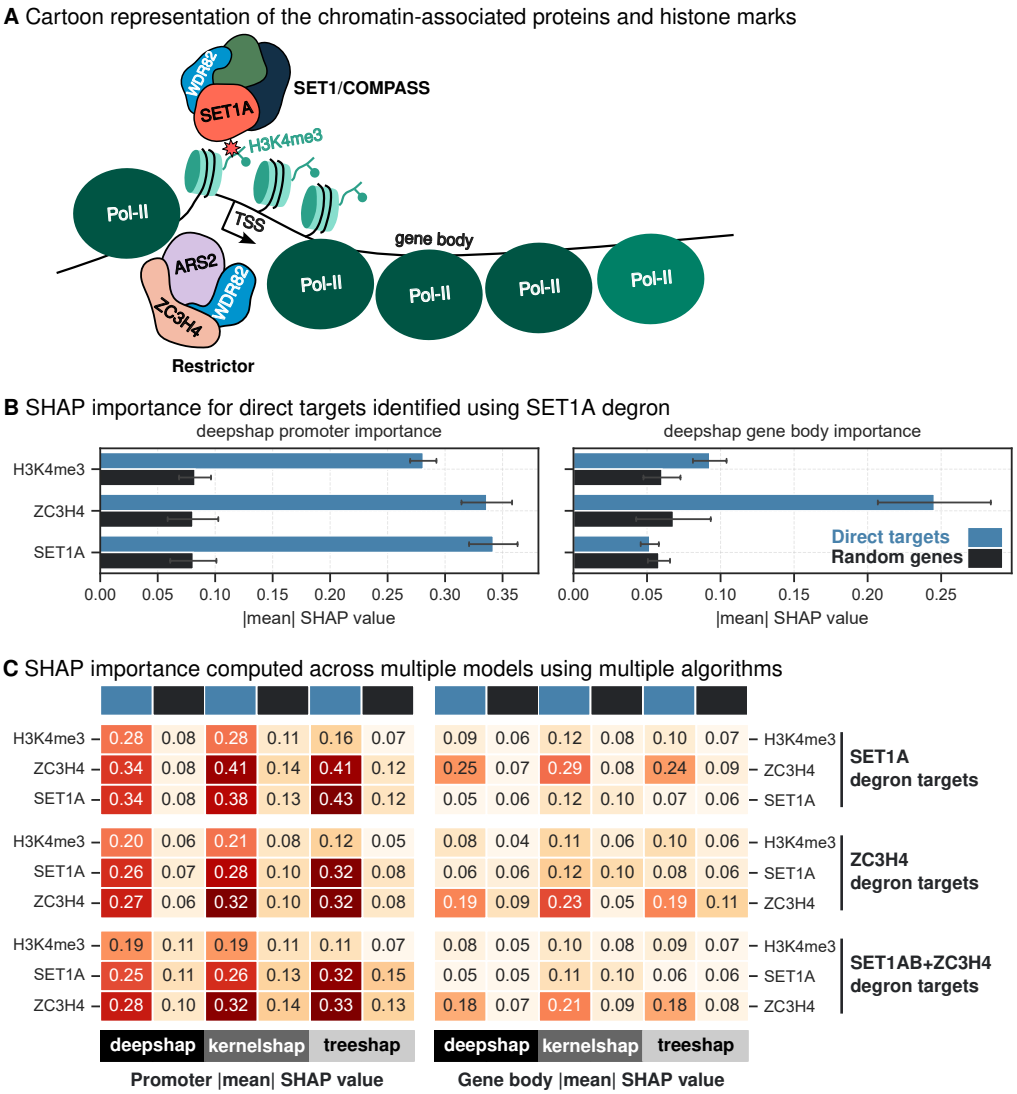


Fig 3. SHAP values for SET1A, H3K4me3 and ZC3H4. (A) Chromatin-associated proteins and histone marks relevant for this model. (B) Comparison of absolute mean DeepSHAP values computed for direct targets (SET1A degron) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. (C) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) following SET1A degradation, ZC3H4 degradataion and SET1A+ZC3H4 degradation simultaneously.

Cross-dataset validation of SHAP analysis suggests interplay between ZC3H4 and INTS11

To investigate the generalizability of our findings across datasets, we performed a cross-dataset comparison by analysing gene sets derived from the degradation of INTS11 (Wang et al. [22]) alongside predictive models trained on independent ChIP-seq data from Hughes et al. [21]. We also conducted the reverse analysis, using gene sets identified as likely direct targets of SET1A, ZC3H4, and the simultaneous degradation of SET1A and ZC3H4 (Hughes et al. [21]) to validate predictive models trained on ChIP-seq data from Wang et al. [22]. This bipartite approach enabled us to evaluate whether SHAP values derived from one dataset could reliably capture the transcriptional regulatory roles in an independent dataset, providing a stringent test of both the predictive model and the SHAP framework. While DPY30 and RBBP5 are core components of the SET1/COMPASS family of complexes [22,27], they

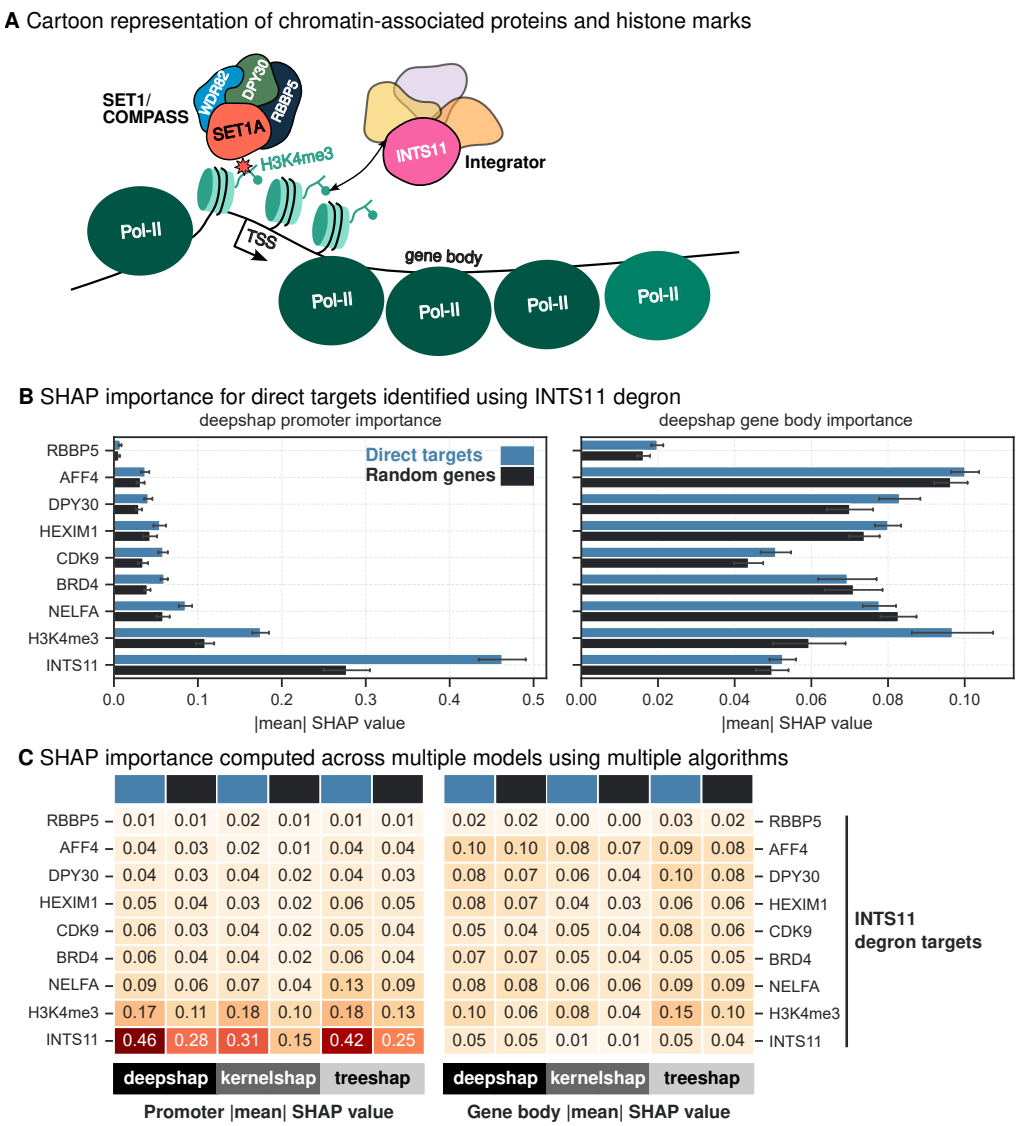


Fig 4. SHAP values for chromatin-associated proteins profiled in Wang et al. [22] (A) Chromatin-associated proteins and histone marks relevant for this model. (B) Comparison of absolute mean DeepSHAP values computed for direct targets (INTS11 degron) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. (C) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) following INTS11 degradation.

are shared across multiple paralogous assemblies [51]. Therefore, we decided to omit DPY30 and RBBP5 degron target genes as they may contain marks deposited by other methyltransferases [52]. Based on the antagonistic relationship between SET1A and ZC3H4 [21] and INTS11's recruitment via H3K4me3 [22], we hypothesise a potential interplay between SET1A, ZC3H4 and INTS11. As predicted, DeepSHAP analysis on Hughes et al. model showed higher SHAP values for SET1A and ZC3H4 at INTS11 target gene loci (S5 FigA). When we computed DeepSHAP values for Wang et al. model, the comparative analysis highlighted INTS11 and H3K4me3 as highly important at SET1A target genes (S5 FigB), ZC3H4 target genes (S5 FigC) and SET1A+ZC3H4 target genes (S5 FigD). These observations were consistent when we extended this analysis using multiple SHAP algorithms (Fig 5A).

To probe this potential connection further, we analysed the overlap of target genes between SET1A, INTS11 and ZC3H4. This analysis revealed that SET1A, INTS11 and ZC3H4 share a significant number of target loci (Fig 5B), with highly correlated transcriptional regulatory patterns between INTS11 and ZC3H4 (Fig 5C) suggesting a functional relationship between these two proteins. Reassuringly, we observed anti-correlation between ZC3H4 and SET1A (Fig 5D) transcriptional changes validating the results in Hughes et al. [21]. Surprisingly, transcriptional regulatory patterns anti-correlated between INTS11 and SET1A (Fig 5E), an observation that stands in contrast to observations in Wang et al. [22]. Existing evidence supports the notion that transcriptional machineries often co-occupy the same loci [49], influencing one another's activity. INTS11 is reported to be recruited by H3K4me3 [22] which could explain this unexpected overlap. This reciprocal analysis and significant overlap between target genes of ZC3H4 and INTS11 provides additional support for our Integrator-Restrictor interplay hypothesis.

To test whether SHAP can evaluate a model generated from multiple studies, we combined ChIP-seq data sets from Hughes and Wang studies [21, 22] to create a unified data set that includes profiles of key regulatory chromatin proteins and modifications, including SET1A, ZC3H4, INTS11 and H3K4me3 (S7 Fig). Model learned from this unified data set was highly accurate (Fig 1B) and this integrative approach allowed us to assess whether SHAP-derived feature importance remains consistent across multiple studies. DeepSHAP analysis consistently identified ZC3H4, SET1A, INTS11 and H3K4me3 promoter occupancy and ZC3H4 gene body occupancy as most important features for target genes identified from INTS11 degron (S6 FigA), SET1A degron (S6 FigB) and ZC3H4 degron (S6 FigC). When we extended our analysis across multiple SHAP algorithms, we reinforced our findings (Fig 6A). This analysis suggest an unexpected interplay between Integrator and Restrictor complexes and reinforces ZC3H4's dual role in transcriptional regulation, functioning both at promoters and along gene bodies.

SHAP importance predicts direct targets and correlates with magnitude of differential gene expression

We next sought to determine whether SHAP importance values could better predict transcriptional changes compared to normalised ChIP signals. To evaluate this, genes present in test data were ranked based on their summed SHAP importance values for highly relevant marks (SET1A, INTS11, ZC3H4) and compared to rankings based on normalised ChIP signal of corresponding proteins for which degron experiments are performed. For every ranked gene by SHAP value or normalised ChIP signal, the fraction of target genes within these ranked genes was calculated. Our analysis revealed that genes ranked by SHAP importance consistently exhibited a higher fraction of differentially expressed genes than those ranked by normalised ChIP signals (Fig 6B). This demonstrates that genes identified by ranking of SHAP values are more likely to be direct targets and these insights can be gained without performing degron experiments. Additionally, SHAP values showed stronger correlations with magnitude of differential gene expression compared to correlations with normalised ChIP signal values (Fig 6C). Ranking of absolute mean SHAP values (e.g. Fig 6A) highlight which chromatin-associated proteins to prioritise in future time-resolved degron studies, and the resulting experiments can validate the functional links predicted by our framework. Overall, our analysis highlight the potential of interpretable modelling and the utility of SHAP for data-driven research, hypothesis testing and validation in the genomics field.

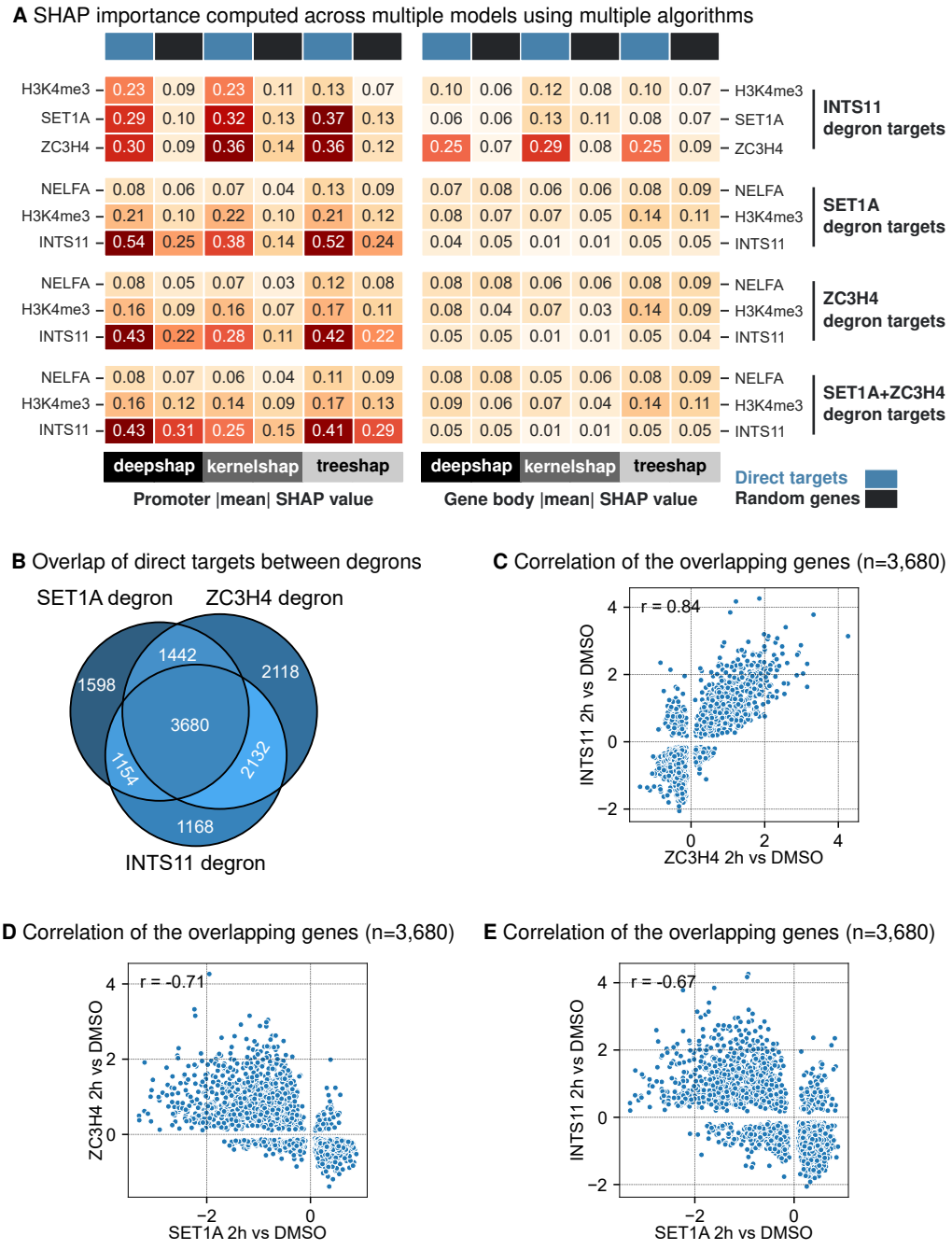
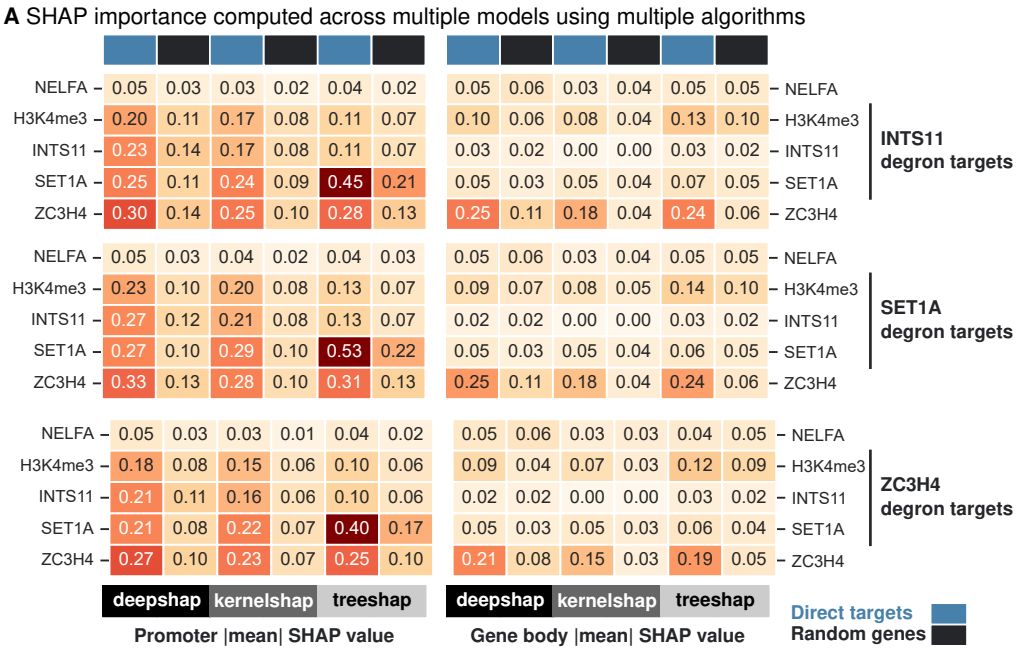
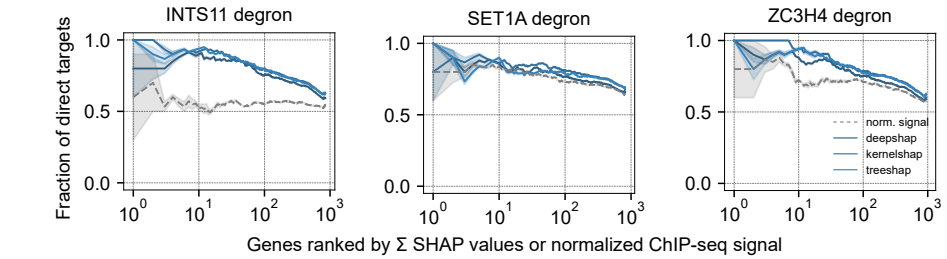


Fig 5. SHAP values for chromatin-associated proteins profiled in Hughes et al. [21] and Wang et al. [22] (A) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) following INTS11 degradation, SET1A degradation, ZC3H4 degradataion and SET1A+ZC3H4 degradation simultaneously. (B) Overlap of direct targets of SET1A, ZC3H4 and INTS11. (C) Scatter plot of log_2 fold changes following ZC3H4 degradation and INTS11 degradation for the overlapping genes. (D) Scatter plot of log_2 fold changes following ZC3H4 degradation and SET1A degradation for the overlapping genes. (E) Scatter plot of log_2 fold changes following SET1A degradation and INTS11 degradation for the overlapping genes.





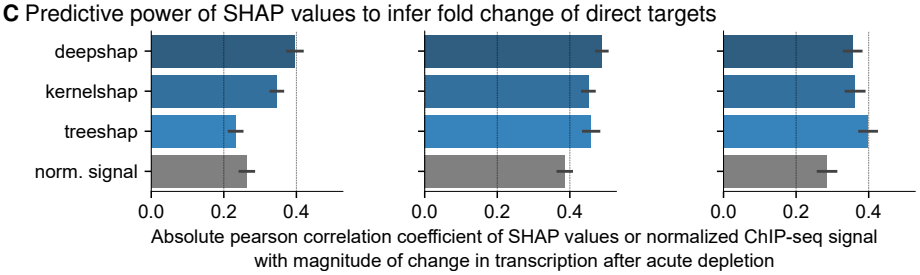


Fig 6. SHAP values for chromatin-associated proteins profiled in integrative model from Hughes et al. [21] and Wang et al. [22] (A) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) following INTS11 degradation, SET1A degradation and ZC3H4 degradataion. (B) Fraction of overlapping direct target genes with genes ranked by summed SHAP values of INTS11, SET1A and ZC3H4 (blue lines) compared to genes ranked by normalised ChIP-seq signal (grey lines). (C) Absolute pearson correlation between SHAP values from different algorithms and log_2 fold change following acute depletion of INTS11, SET1A and ZC3H4. As a control comparison, normalised ChIP-seq was correlated with log_2 fold change. Error is estimated by calculating correlation coefficients for all 5 splits and the error bars indicate the standard deviation of one-half.

Discussion

A key strength of this study lies in the ability of SHAP analysis to identify biologically meaningful relationships directly from data derived from unperturbed cell systems. The model, trained solely on unperturbed conditions, captures the relationships between protein occupancy and transcriptional output without prior knowledge of perturbations. Remarkably, by ranking genes based on SHAP importance values, direct targets of perturbation can be accurately predicted. This represents a potentially significant achievement, enabling context-dependent inference and complex interplay of regulatory targets without the need for time-consuming and expensive perturbation experiments. We also evaluated differences in SHAP-derived importance for target gene sets derived from perturbation experiments to infer regulatory roles of individual proteins. For instance, the higher SHAP importance of ZC3H4 at promoters and gene bodies of target genes suggests distinct regulatory roles for ZC3H4 and SET1A, highlighting how the model trained on unperturbed systems can reveal insights under perturbed conditions. This observation aligns with recent experimental evidence that ZC3H4 recruits the PNUTS phosphatase [53], and directly controls the phosphorylation of proteins involved in regulating pause release and elongation [54,55], providing further insight into ZC3H4-mediated transcriptional regulation. In parallel, the consistent importance of INTS11 promoter occupancy highlights its critical role in transcriptional regulation via H3K4me3 recruitment [22]. Cross-dataset validation and integrative modelling reinforced the utility of SHAP-based interpretations, revealing unexpected overlaps between ZC3H4 and INTS11 target loci. This functional connection, supported by shared transcriptional regulatory patterns, raises the intriguing possibility of cooperation between the Restrictor and Integrator complexes (S7 Fig), potentially converging on a common transcriptional termination pathway [49]. Whether the recruitment of INTS11 via H3K4me3 [22] or the antagonistic interplay between ZC3H4 and SET1A [21], which deposits H3K4me3, stimulates a cascade of transcriptional regulation at the same genes remains an open question. Future experiments will be essential to unravel the precise mechanisms underlying these interactions and to determine whether they act sequentially, interdependently, or as part of a larger coordinated regulatory pathway.

The analysis of PRC1 and PRC2 regulatory systems highlights both the strengths and challenges of SHAP in disentangling contributions of co-occupying proteins. SHAP consistently identified RING1B, the core PRC1 subunit, as a key contributor to transcriptional repression, aligning with research that established PRC1's dominant role at early time points [20,38,39]. This result demonstrates SHAP's robustness in capturing biologically meaningful relationships, even in highly co-linear systems. The predominant contribution of EZH2 at some time points signifies the complex interplay between PRC1 and PRC2 [25,41–43]. Notably, SHAP's ability to identify H3K4me3 importance at polycomb-repressed loci suggests potential recruitment of the SET1 complex, uncovering new avenues for understanding the interplay between activation and repression machinery [46]. It could also indicate identification of "bivalent" promoters [47,48], which are a peculiar feature of embryonic stem cells. Overall, these results reaffirm SHAP's utility as a powerful tool for identifying key regulators and generating novel biological hypotheses, while acknowledging the challenges posed by systems with overlapping regulatory inputs.

Several papers have aimed to disentangle the complex network of biochemical signals underpinning the initiation of transcription [56,57]. However, most of these approaches proceeded from a simplifying hypothesis that the expression of different genes are independent processes, conditioned on the local epigenetic environment. Such an assumption necessarily can only infer an average effect of a specific epigenetic factor on gene expression, and requires a very comprehensive assaying of epigenetic marks to be approximately valid. A major benefit of our method is that SHAP values are per gene, and thus directly provide an estimate of context dependence which is increasingly recognised to be a necessary ingredient to understand the mechanisms of gene expression [58]. Our case study highlights the significant advantages of SHAP in providing gene-by-gene explanations for individual predictions, enabling insights into transcriptional regulation that would be difficult to achieve using traditional methods. By integrating predictive modelling with explainable AI, we offer a scalable framework capable of exploring both well-characterised and poorly understood transcriptional processes. A known limitation of SHAP values calculated using the KernelSHAP algorithm is the implicit assumption that input features are statistically independent. For epigenomic data this assumption is rarely met because of the co-localisation of histone modifications, chromatin-associated proteins and RNA Pol-II. We therefore employed other algorithms like TreeSHAP and DeepSHAP which attempt to account for some statistical dependence. We emphasise that SHAP values, regardless of the algorithm used, quantify associations and does not direct causality. We recommend they should always be interpreted in combination with experimental evidence. While the present model centres on promoter- and gene body- occupancies, incorporating regulation through enhancers

could broaden its predictive scope, though reliably mapping enhancers to their target genes is an emerging challenge. Overall, this study underscores the transformative potential of combining computational modelling with explainable AI to advance our understanding of transcriptional regulation and generate novel biological hypotheses.

Materials and methods

Preprocessing of sequencing data

All bioinformatic analyses were performed using Snakemake workflows [59], ensuring reproducibility and scalability of the analysis pipeline explained in Supplementary Methods. These workflows include detailed specifications of software versions, command-line parameters, and dependencies for each analysis step.

ChIP-seq data analysis

Raw sequencing data for ChIP-seq experiments were downloaded from the Gene Expression Omnibus (GEO). Reads were aligned to the reference genome using Bowtie2 [60], ensuring high alignment accuracy. To improve data quality, reads mapping to ENCODE blacklisted regions, which are repetitive and highly mappable, were removed. Peak calling and signal track generation were performed with MACS2 [61]. Treatment and control pileup tracks were generated, and the `bdgcmp` command was used to calculate fold enrichment tracks, normalising ChIP signal to input samples. To quantify the signal, bigWigAverageOverBed [62] was used to calculate the mean signal across defined genomic regions, such as promoters and gene bodies. For each ChIP-seq feature, the average signal intensities were calculated across gene promoters (±1 kb from transcription start site, TSS) and gene bodies (+1kb from TSS to transcription end site, TES). These quantifications were performed using bigWigAverageOverBed, which computes the average signal intensity over specified genomic intervals.

RNA-seq and TT-seq data analysis

Raw sequencing data for RNA-seq and TT-seq experiments were also downloaded from GEO. Reads were aligned to the reference genome using Bowtie2 [60]. PCR duplicates were identified and removed using SAMtools markup [63] to eliminate artifacts due to library preparation. Gene-level read counts were quantified using featureCounts [64] from the Subread package. Differential gene expression analysis was performed using DESeq2 [65] in R, applying default parameters to identify significantly differentially expressed genes.

Sequencing data sets

Data sets considered for interpretable modelling are listed in Table 1. These data sets include ChIP-seq and RNA-seq data from mouse embryonic stem cells (mESCs) with perturbations targeting chromatin-associated proteins and histone modifications. The data sets were selected based on their relevance to transcriptional regulation and the availability of high-quality sequencing data.

Data preparation and modelling

To perform SHAP-based model interpretation, we first processed publicly available chromatin profiling ChIP-seq data. Raw sequencing reads and corresponding input controls were downloaded from GEO. All data sources, associated GEO accessions, and chromatin-associated proteins used in this study are listed in Table 1. All reads were processed using a standardised pipeline as described in ChIP-seq data analysis, detailed and its steps visualised in Supplementary Methods. The processed data were then used to generate a feature matrix, where each row corresponds to a gene and each column corresponds to a chromatin-associated protein or histone modification. The feature matrix was constructed by averaging the ChIP-seq signal over promoter regions (±1 kb from TSS) and gene bodies (+1 kb from TSS to TES) for each gene.

To predict RNA polymerase II occupancy and assess the contribution of chromatin features, we first generated five distinct train-validation-test (80-10-10%) splits of the dataset. Following this, the data was log_{10} transformed, then

Table 1. Sequencing data sets considered for interpretable modelling

Accession	Source	Transcription data from degrons (RNA-seq and TT-chem-seq)	Chromatin associated proteins and histone modifications (ChIP-seq)
GSE199805 [21]	mESC	SET1A dTAG, ZC3H4 dTAG, SET1A+ZC3H4 dTAG	SET1A, ZC3H4, H3K4me3, Pol-II
GSE181714 [22]	mESC	INTS11 dTAG	INTS11, DPY30, NELFA, H3K4me3, CDK9, AFF4, HEXIM1, BRD4, RBBP5, Pol-II
GSE199805 [21], GSE181714 [22]	mESC	INTS11 dTAG, SET1A dTAG, ZC3H4 dTAG, SET1A+ZC3H4 dTAG	INTS11, DPY30, NELFA, H3K4me3, CDK9, AFF4, HEXIM1, BRD4, RBBP5, SET1A, ZC3H4, Pol-II
GSE159400 [20]	mESC	RING1B dTAG	RING1B, SUZ12, H2AK119ub1, H3K27me3, H3K4me3, Pol-II

centered to zero mean and unit variance. Each of the four model types was independently trained on all five splits to evaluate performance variability and robustness. The models included: (1) a feedforward neural network implemented in PyTorch [66], with hyperparameters (learning rate initiation, weight decay and hidden layer sizes) with ReLu as activation function optimised using Optuna [67]; (2) an MLPRegressor from scikit-learn [68] with hyperparameters (L2 regularization alpha, hidden layer sizes, and a choice of activation function); (3) an XGBoostRegressor from the XGBoost package [69], tuned for learning rate (1e-3 to 1e-1), tree_method ("exact" or "hist"), and max_depth (6 to 18); and (4) a baseline LinearRegression model from scikit-learn [68]. Before commencing training, the random seed was set to 42. Training, validation and test R^2 scores were recorded for all models across the five data splits. Performance was evaluated by plotting R^2 values for training, validation, and test sets for both output targets across all configurations, as shown in Fig 1B and S2 FigA. The hyperparameters learned from the training and validation sets are tabulated and the training and validation R^2 history for each model across the five splits for every epoch is shown in Supplementary Methods.

Prediction and interpretability analysis

All models were trained to output two distinct values per gene: RNA polymerase II (Pol-II) signal at the promoter and Pol-II signal across the gene body. Accordingly, SHAP values were computed separately for each output using model-appropriate explainers: shap.DeepExplainer for the PyTorch model, shap.KernelExplainer for the MLPRegressor and shap.TreeExplainer for the XGBoostRegressor. For the baseline LinearRegression model, 95% confidence intervals for regression coefficients were calculated. For each gene, SHAP values for promoter features (i.e., chromatin features summarised over promoter regions) were calculated with respect to the Pol-II promoter signal, while SHAP values for gene body features were calculated with respect to the Pol-II gene body signal. This ensured that feature importance was interpreted in the appropriate spatial and functional context for each output.

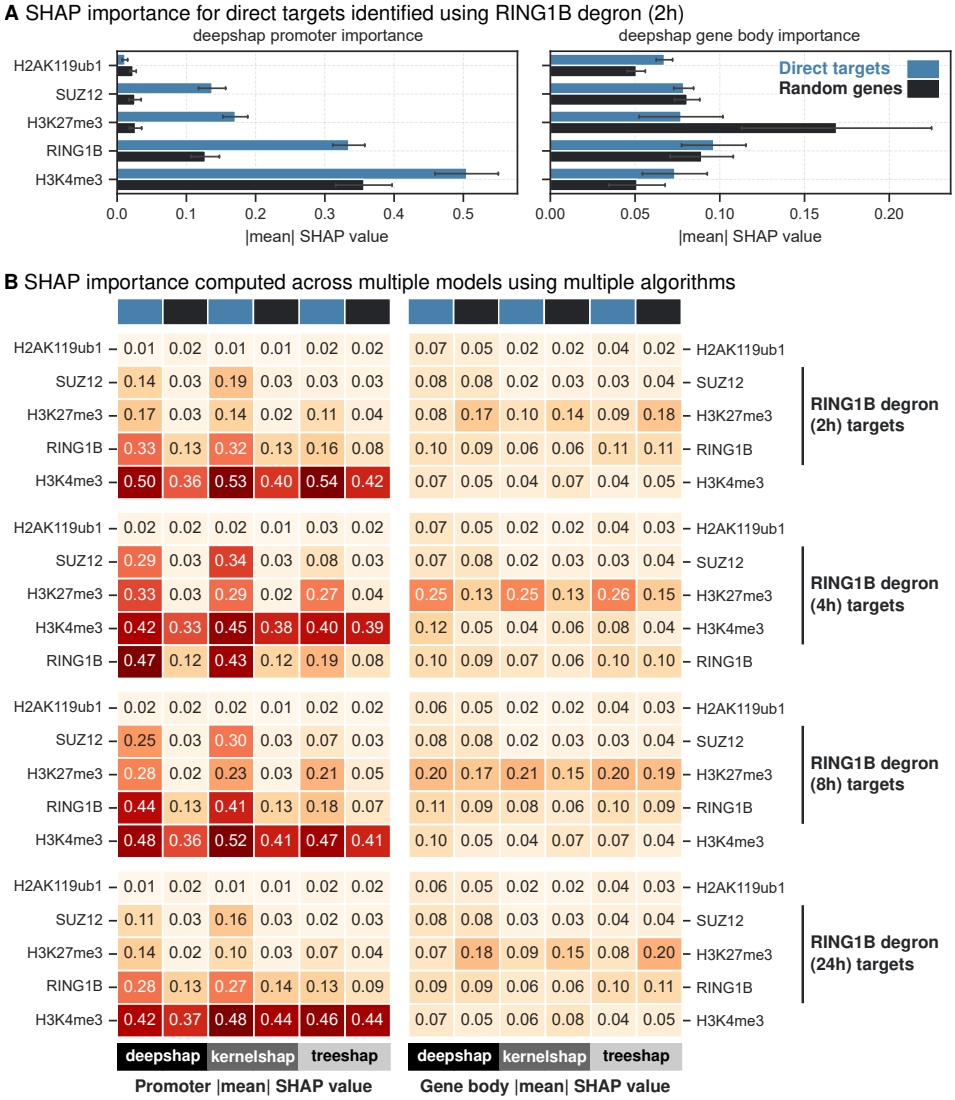
The background dataset for SHAP analysis was generated using shap.kmeans, excluding any genes identified as direct targets in the corresponding TT-seq perturbation to avoid information leakage. To contextualize SHAP values with transcriptional response, we stratified test-set genes using adjusted p-value thresholds (below and above 0.05) from TT-seq differential expression experiments. For each threshold category, 50 genes were randomly sampled and their absolute mean SHAP values were computed across 50 iterations to evaluate intra-category variability.

We computed pairwise Pearson correlations for every dataset and visualised them as clustered heat-maps, which revealed substantial positive correlations and a clear separation between promoter and gene body signals. Despite this multicollinearity, the mean absolute SHAP scores and the rank order of feature importance were essentially identical across DeepSHAP, KernelSHAP and TreeSHAP (and across train/test splits), demonstrating that our interpretability results are robust to shared signal among features. This analysis is shown in Supplementary Methods.

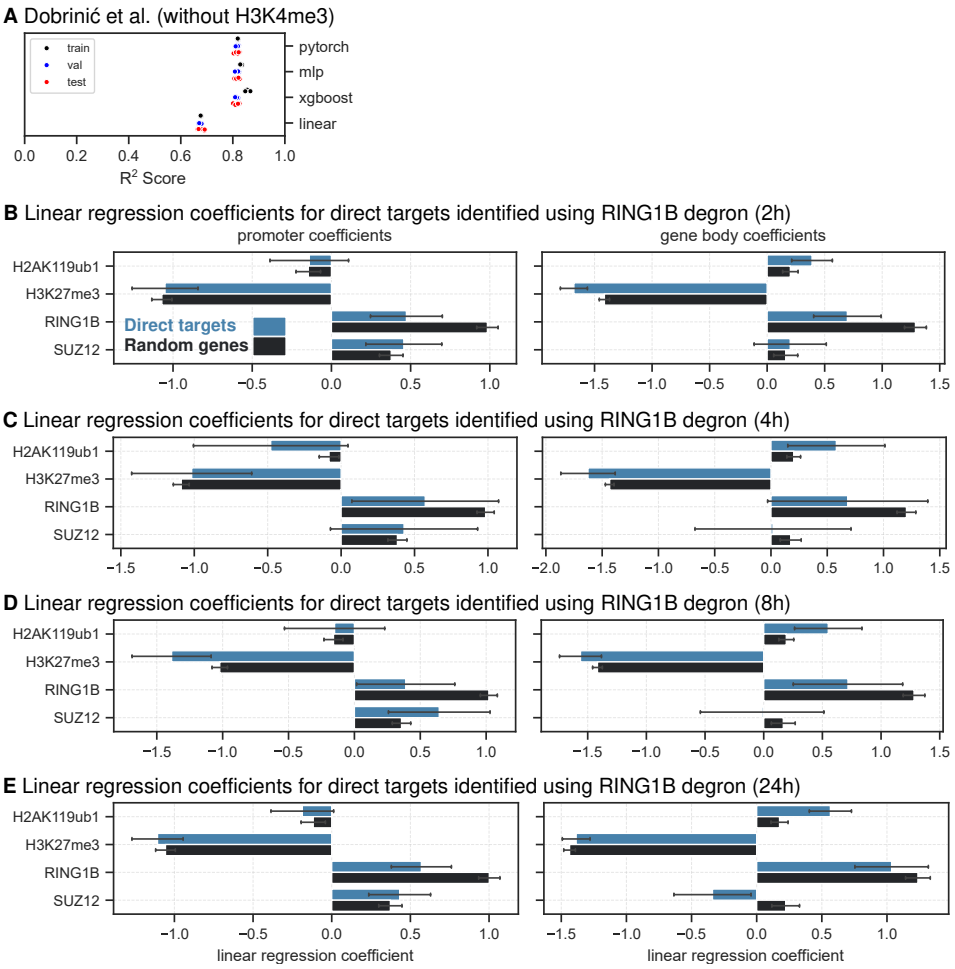
Software and code availability

All modelling analyses, including data preparation, MLP regression [66], XGBoost Regression [69], SHAP analysis [17], and linear regression [68], were performed using Python. The source code is publicly available on the GitHub repository https://github.com/kashyapchhatbar/SHAP-analysis.

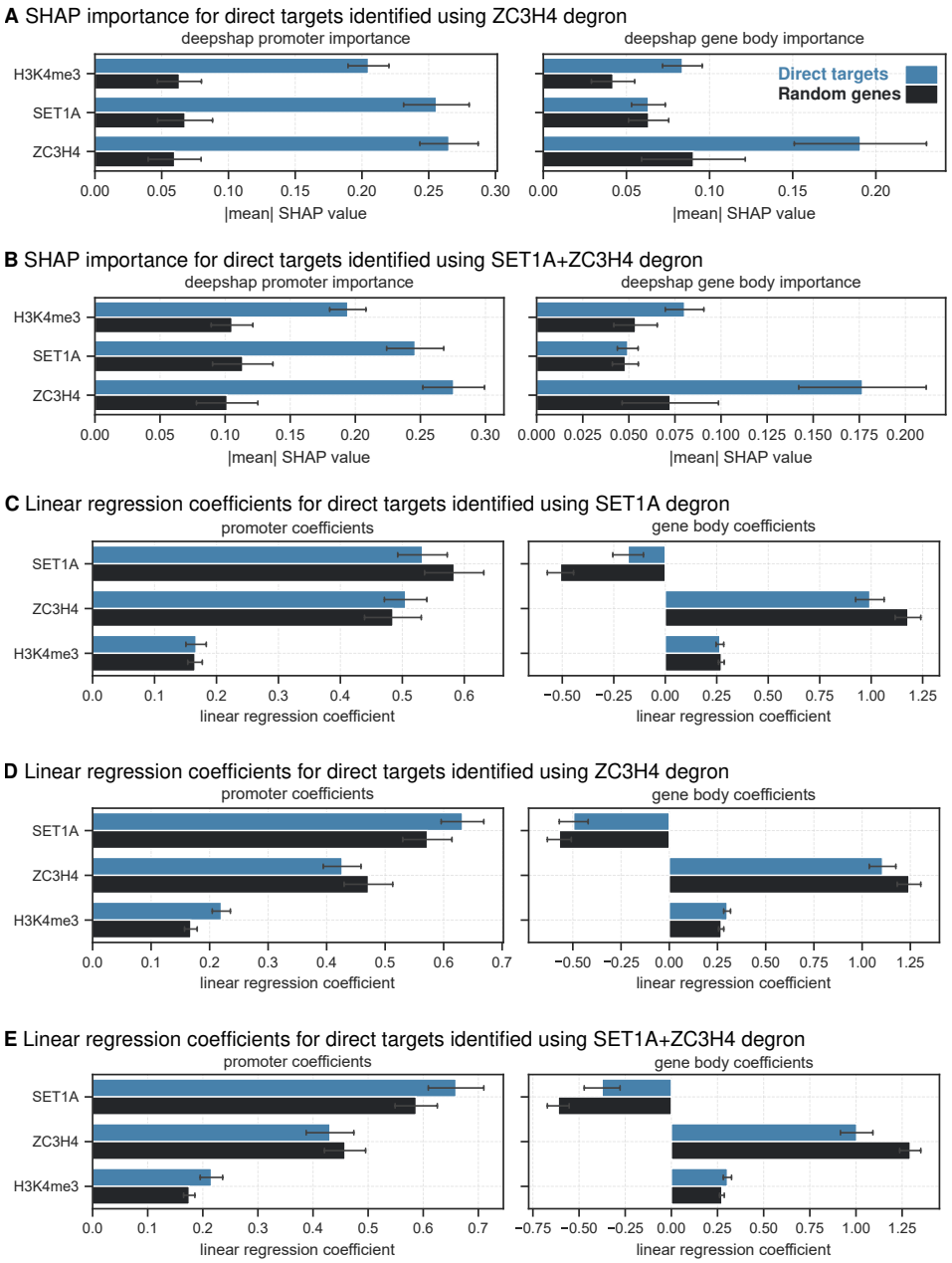
Supporting information



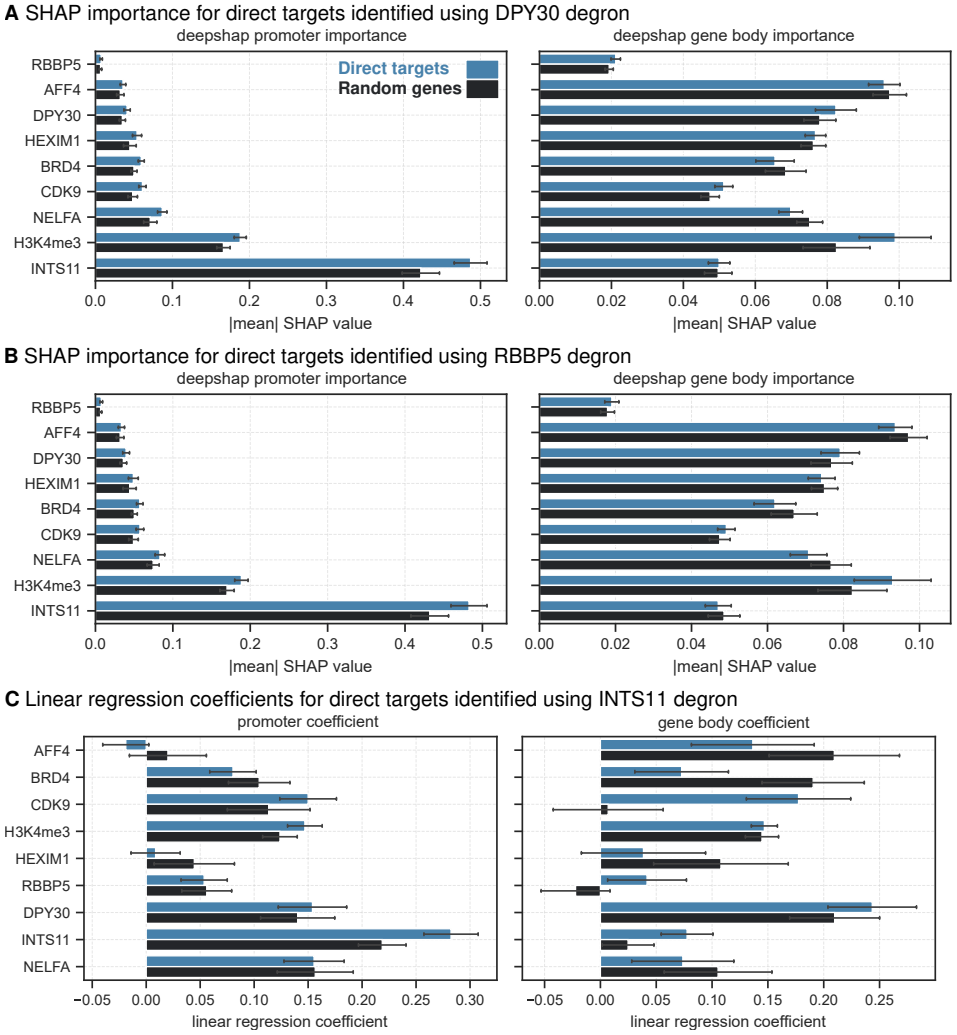
S1 Fig. SHAP values for H3K4me3, co-occupying PRC1 (RING1B) and PRC2 (SUZ12) components and their catalytic marks across multiple time points following RING1B degradation. (A) Absolute mean DeepSHAP values computed for direct targets (RING1B degron 2h) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. (B) Heatmap of absolute mean SHAP values for different algorithms DeepSHAP, KernelSHAP and TreeSHAP is plotted for direct targets (blue) and random genes (black) across all time points following RING1B degradation (2h, 4h, 8h and 24h).



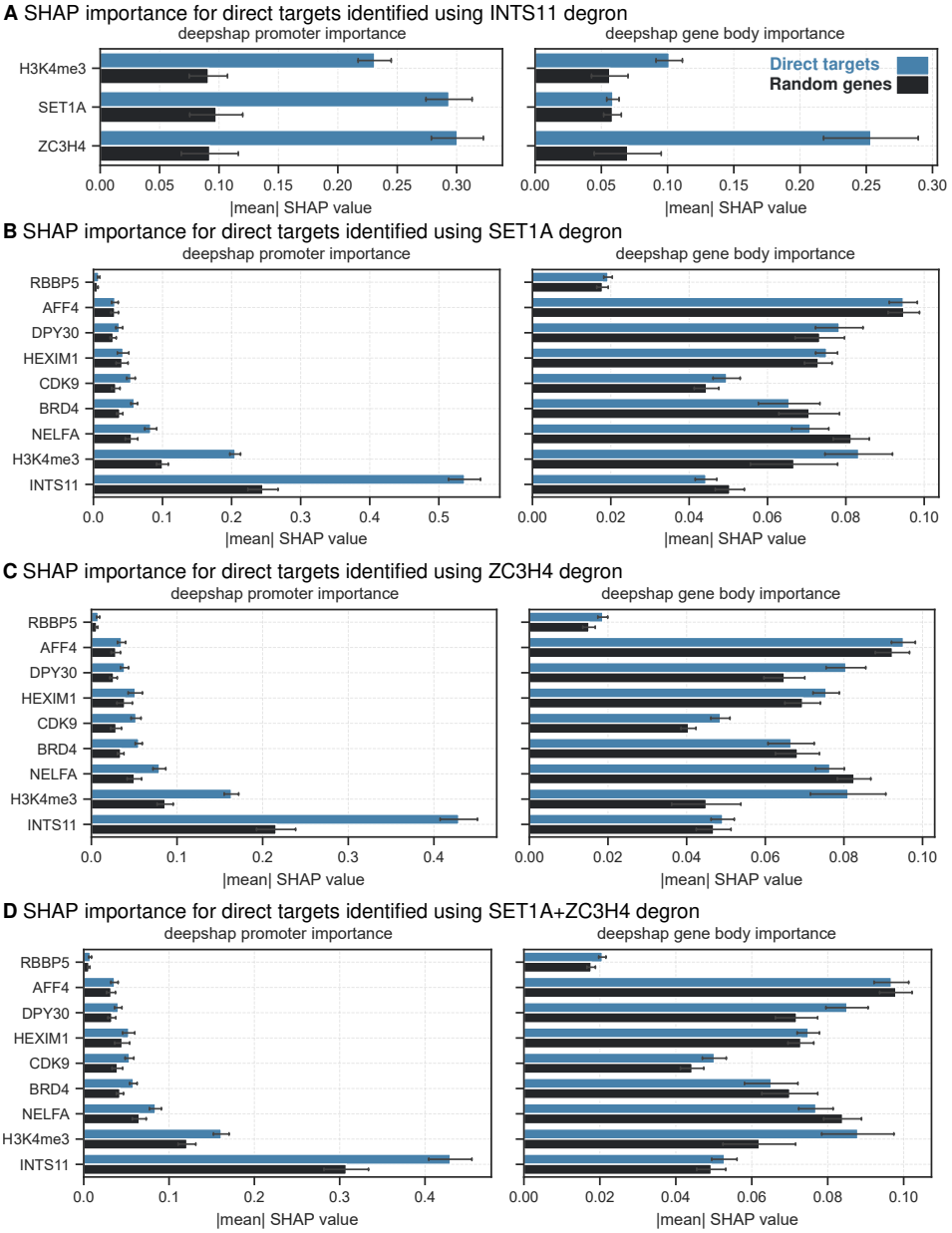
S2 Fig. Model performance and linear regression coefficients for co-occupying PRC1 (RING1B) and PRC2 (SUZ12) components and their catalytic marks across multiple time points following RING1B degradation. (A) Model performance for predicting RNA Pol-II occupancy based on chromatin-associated protein profiles for Dobrinić et al. without incorporating H3K4me3. Coefficients of determination (R^2) are shown for training, validation and test datasets for 5 splits. Coefficients for chromatin-associated proteins and histone marks for direct targets (blue) and random genes (black) for (B) 2h (C) 4h (D) 8h and (E) 24h following RING1B degradation.



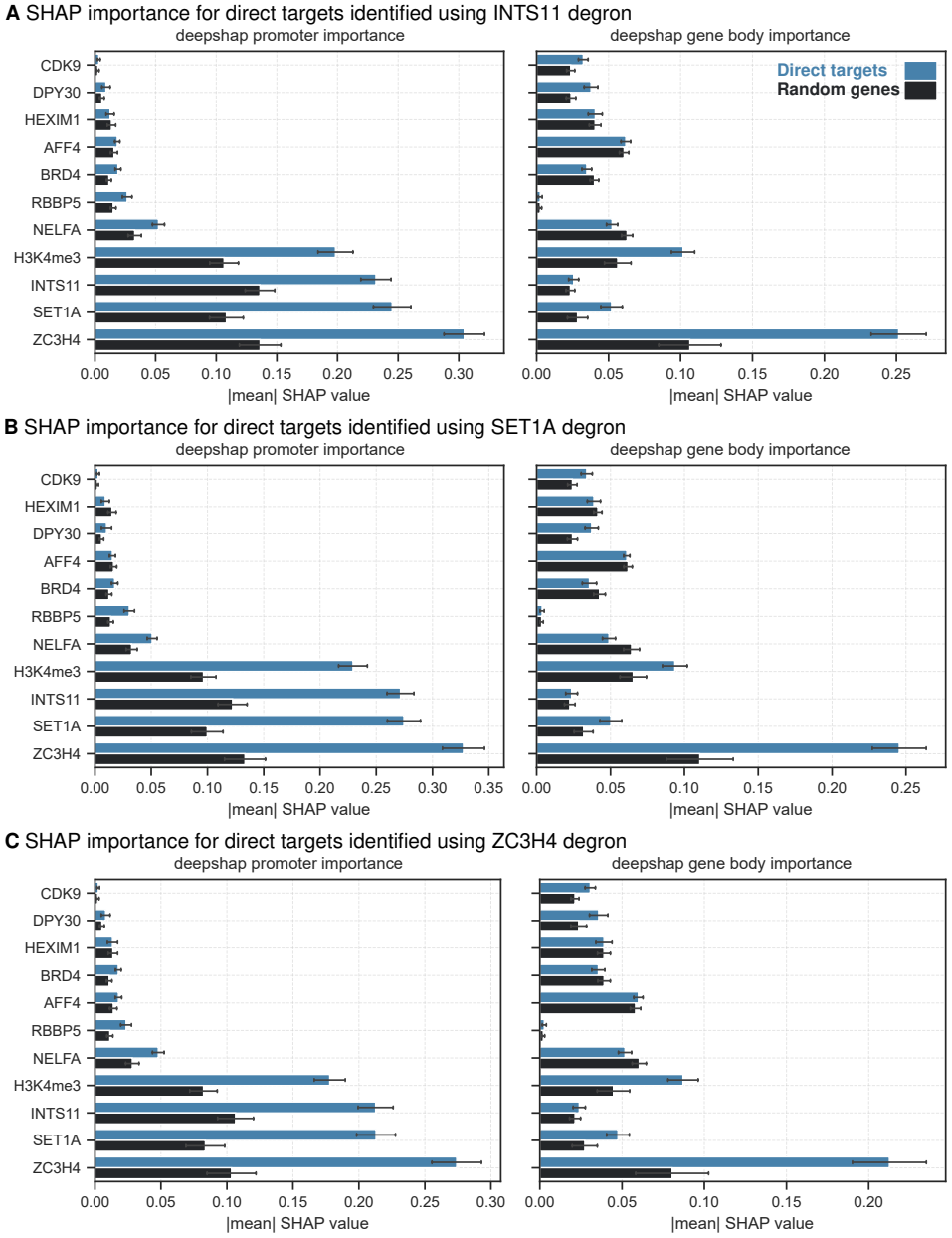
S3 Fig. SHAP values and linear regression coefficients for SET1A, H3K4me3 and ZC3H4. Comparison of absolute mean DeepSHAP values computed for (A) direct targets (ZC3H4 degron) and random genes, (B) direct targets (SET1A+ZC3H4 degron) at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. Coefficients for chromatin-associated proteins and histone marks for direct targets (blue) and random genes (black) for (C) SET1A degron (D) ZC3H4 degron and (E) SET1A+ZC3H4 degron.



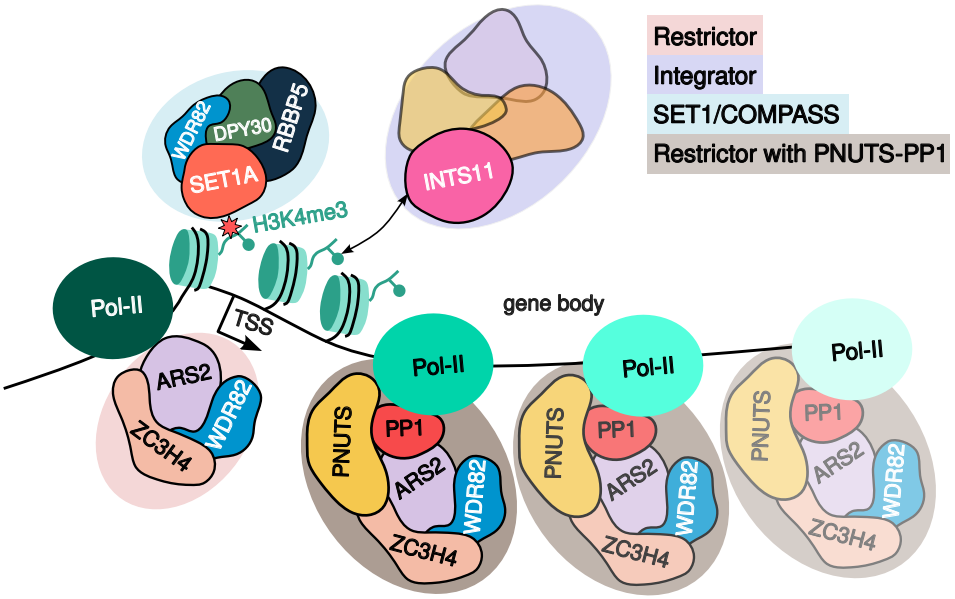
S4 Fig. SHAP values and linear regression coefficients for chromatin-associated proteins profiled in Wang et al. [22]. Comparison of absolute mean DeepSHAP values computed for (A) direct targets (DPY30 degron) and random genes, (B) direct targets (RBBP5 degron) at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half. Coefficients for chromatin-associated proteins and histone marks for direct targets (blue) and random genes (black) for (C) SET1A degron (D) ZC3H4 degron and (E) SET1A+ZC3H4 degron.



S5 Fig. SHAP values for chormatin-associated proteins profiled in Hughes et al. [21] and Wang et al. [22] Absolute mean DeepSHAP values computed for (A) direct targets (INTS11 degron) (B) direct targets (SET1A degron) (C) direct targets (ZC3H4 degron) (D) direct targets (SET1A+ZC3H4 degron) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half.



S6 Fig. SHAP values for chormatin-associated proteins profiled in Hughes et al. [21] and Wang et al. [22] from the integrated model. Absolute mean DeepSHAP values computed for (A) direct targets (INTS11 degron) (B) direct targets (SET1A degron) (C) direct targets (ZC3H4 degron) and random genes at promoter and gene body contexts. Random subsampling is used to calculate intra-category variability and the error bars indicate the standard deviation of one-half.



S7 Fig. Potential crosstalk between Integrator and Restrictor complexes in transcriptional regulation

Acknowledgments

We are grateful to Sara Giuliani for helpful discussions and feedback on the manuscript.

References

1. Ptashne M, Gann A. Imposing specificity by localization: mechanism and evolvability;8(22):R812–R822. doi:10.1016/S0960-9822(07)00508-8.
2. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome;489(7414):57–74. doi:10.1038/nature11247.
3. Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes;98(10):1487–1489. doi:10.3324/haematol.2013.094243.
4. Stunnenberg HG, Abrignani S, Adams D, Almeida Md, Altucci L, Amin V, et al. The international human epigenome consortium: a blueprint for scientific collaboration and discovery;167(5):1145–1149. doi:10.1016/j.cell.2016.11.007.
5. Benveniste D, Sonntag HJ, Sanguinetti G, Sproul D. Transcription factor binding predicts histone modifications in human cell lines;111(37):13367–13372. doi:10.1073/pnas.1412081111.
6. Beer MA, Tavazoie S. Predicting gene expression from sequence;117(2):185–198. doi:10.1016/S0092-8674(04)00304-6.
7. Budden DM, Hurley DG, Crampin EJ. Predictive modelling of gene expression from transcriptional regulatory elements;16(4):616–628. doi:10.1093/bib/bbu034.
8. Wilczynski B, Liu YH, Yeo ZX, Furlong EEM. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state;8(12):e1002798. doi:10.1371/journal.pcbi.1002798.

9. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells;106(51):21521–21526. doi:10.1073/pnas.0904863106.

10. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells;40(2):553–568. doi:10.1093/nar/gkr752.

11. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets;12(2):R15. doi:10.1186/gb-2011-12-2-r15.

12. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning;18(1):67. doi:10.1186/s13059-017-1189-z.

13. Schreiber J, Durham T, Bilmes J, Noble WS. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome;21(1):81. doi:10.1186/s13059-020-01977-6.

14. Durham TJ, Libbrecht MW, Howbert JJ, Bilmes J, Noble WS. PREDICTD PaRallel epigenomics data imputation with cloud-based tensor decomposition;9(1):1402. doi:10.1038/s41467-018-03635-9.

15. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues;33(4):364–376. doi:10.1038/nbt.3157.

16. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16. Association for Computing Machinery;. p. 1135–1144. Available from: https://dl.acm.org/doi/10.1145/2939672.2939778.

17. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. vol. 30. Curran Associates, Inc.;.Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

18. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: An overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Explainable AI: Interpreting, explaining and visualizing deep learning. Springer International Publishing;. p. 193–209. Available from: https://doi.org/10.1007/978-3-030-28954-6_10.

19. Jaeger MG, Winter GE. Fast-acting chemical tools to delineate causality in transcriptional control;81(8):1617–1630. doi:10.1016/j.molcel.2021.02.015.

20. Dobrinić P, Szczurek AT, Klose RJ. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency;28(10):811–824. doi:10.1038/s41594-021-00661-y.

21. Hughes AL, Szczurek AT, Kelley JR, Lastuvkova A, Turberfield AH, Dimitrova E, et al. A CpG island-encoded mechanism protects genes from premature transcription termination;14(1):726. doi:10.1038/s41467-023-36236-2.

22. Wang H, Fan Z, Shliaha PV, Miele M, Hendrickson RC, Jiang X, et al. H3K4me3 regulates RNA polymerase II promoter-proximal pause-release;615(7951):339–348. doi:10.1038/s41586-023-05780-8.

23. de Napoles M, Mermoud JE, Wakao R, Tang YA, Endoh M, Appanah R, et al. Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation;7(5):663–676. doi:10.1016/j.devcel.2004.10.005.

24. Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS, et al. Role of histone H2A ubiquitination in Polycomb silencing;431(7010):873–878. doi:10.1038/nature02985.

25. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, et al. Role of histone H3 lysine 27 methylation in polycomb-group silencing;298(5595):1039–1043. doi:10.1126/science.1076997.

26. Müller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, et al. Histone methyltransferase activity of a *Drosophila* polycomb group repressor complex;111(2):197–208. doi:10.1016/S0092-8674(02)00976-5.

27. Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, et al. COMPASS: A complex of proteins associated with a trithorax-related SET domain protein;98(23):12902–12907. doi:10.1073/pnas.231473398.

28. Estell C, Davidson L, Steketee PC, Monier A, West S. ZC3H4 restricts non-coding transcription in human cells;10:e67305. doi:10.7554/eLife.67305.

29. Baillat D, Hakimi MA, Näär AM, Shilatifard A, Cooch N, Shiekhattar R. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the c-terminal repeat of RNA polymerase II;123(2):265–276. doi:10.1016/j.cell.2005.08.019.

30. Stein CB, Field AR, Mimoso CA, Zhao C, Huang KL, Wagner EJ, et al. Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci;82(22):4232–4245.e11. doi:10.1016/j.molcel.2022.10.004.

31. Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation;6:e29736. doi:10.7554/eLife.29736.

32. Wu SY, Chiang CM. The double bromodomain-containing chromatin adaptor brd4 and transcriptional regulation*;282(18):13141–13145. doi:10.1074/jbc.R700001200.

33. Li J, Liu Y, Rhee HS, Ghosh SKB, Bai L, Pugh BF, et al. Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing;50(5):711–722. doi:10.1016/j.molcel.2013.05.016.

34. Kwak H, Lis JT. Control of transcriptional elongation;47:483–508. doi:10.1146/annurev-genet-110711-155440.

35. Tong WL, Pehlevan C. MLPs learn in-context on regression and classification tasks;. Available from: http://arxiv.org/abs/2405.15618.

36. Liu H, Dai Z, So DR, Le QV. Pay attention to MLPs;. Available from: http://arxiv.org/abs/2105.08050.

37. Shapley LS. 17. A Value for n-Person Games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press;. p. 307–318. Available from: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html.

38. Blackledge NP, Fursova NA, Kelley JR, Huseyin MK, Feldmann A, Klose RJ. PRC1 catalytic activity is central to polycomb system function;77(4):857–874.e9. doi:10.1016/j.molcel.2019.12.001.

39. Blackledge NP, Klose RJ. The molecular principles of gene regulation by Polycomb repressive complexes;22(12):815–833. doi:10.1038/s41580-021-00398-y.

40. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions;20(9):1123–1136. doi:10.1101/gad.381706.

41. Blackledge NP, Farcas AM, Kondo T, King HW, McGouran JF, Hanssen LLP, et al. Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation;157(6):1445–1459. doi:10.1016/j.cell.2014.05.004.

42. Min J, Zhang Y, Xu RM. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27;17(15):1823–1828. doi:10.1101/gad.269603.

43. Wang L, Brown JL, Cao R, Zhang Y, Kassis JA, Jones RS. Hierarchical recruitment of polycomb group silencing complexes;14(5):637–646. doi:10.1016/j.molcel.2004.05.009.

44. Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, et al. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells;26(15):1714–1728. doi:10.1101/gad.194209.112.

45. Farcas AM, Blackledge NP, Sudbery I, Long HK, McGouran JF, Rose NR, et al. KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands;1:e00205. doi:10.7554/eLife.00205.

46. Klose RJ, Cooper S, Farcas AM, Blackledge NP, Brockdorff N. Chromatin sampling—an emerging perspective on targeting polycomb repressor proteins;9(8):e1003717. doi:10.1371/journal.pgen.1003717.

47. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells;125(2):315–326. doi:10.1016/j.cell.2006.02.041.

48. Voigt P, Tee WW, Reinberg D. A double take on bivalent promoters;27(12):1318–1338. doi:10.1101/gad.219626.113.

49. Estell C, Davidson L, Eaton JD, Kimura H, Gold VAM, West S. A restrictor complex of ZC3H4, WDR82, and ARS2 integrates with PNUTS to control unproductive transcription;83(13):2222–2239.e5. doi:10.1016/j.molcel.2023.05.029.

50. Chen R, Yik JHN, Lew QJ, Chao SH. Brd4 and HEXIM1: Multiple roles in p-TEFb regulation and cancer;2014(1):232870. doi:10.1155/2014/232870.

51. van Nuland R, S Arne H, P Paschalina, J Pascal W T C, V Michiel, , et al. Quantitative Dissection and Stoichiometry Determination of the Human SET1/MLL Histone Methyltransferase Complexes;33(10):2067–2077. doi:10.1128/MCB.01742-12.

52. Piunti A, Shilatifard A. Epigenetic balance of gene expression by Polycomb and COMPASS families;352(6290):aad9780. doi:10.1126/science.aad9780.

53. Erickson B, Fedoryshchak R, Fong N, Sheridan R, Larson KY, Saviola AJ, et al.. PP1 PNUTS binds the "restrictor" and dephosphorylates RNA pol II CTD Ser5 to stimulate transcription termination;. Available from: https://www.biorxiv.org/content/10.1101/2024.07.12.603302v1.

54. Kelley JR, Dimitrova E, Maciuszek M, Nguyen HT, Szczurek AT, Hughes AL, et al. The PNUTS phosphatase complex controls transcription pause release;84(24):4843–4861.e8. doi:10.1016/j.molcel.2024.10.045.

55. Wang Z, Song A, Tao B, Miao M, Luo YQ, Wang J, et al. The phosphatase PP1 sustains global transcription by promoting RNA polymerase II pause release;84(24):4824–4842.e7. doi:10.1016/j.molcel.2024.10.046.

56. Perner J, Lasserre J, Kinkley S, Vingron M, Chung HR. Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling;42(22):13689–13695. doi:10.1093/nar/gku1234.

57. Juan D, Perner J, Carrillo de Santa Pau E, Marsili S, Ochoa D, Chung HR, et al. Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs;14(5):1246–1257. doi:10.1016/j.celrep.2016.01.008.

58. Papp B, Oliver S. Genome-wide analysis of the context-dependence of regulatory networks;6(2):206. doi:10.1186/gb-2005-6-2-206.

59. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine;28(19):2520–2522. doi:10.1093/bioinformatics/bts480.

60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2;9(4):357–359. doi:10.1038/nmeth.1923.

61. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS);9(9):R137. doi:10.1186/gb-2008-9-9-r137.

62. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets;26(17):2204–2207. doi:10.1093/bioinformatics/btq351.

63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

64. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features;30(7):923–930. doi:10.1093/bioinformatics/btt656.

65. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2;15(12):550. doi:10.1186/s13059-014-0550-8.

66. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al.. PyTorch: An Imperative Style, High-Performance Deep Learning Library;. Available from: http://arxiv.org/abs/1912.01703.

67. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework;. Available from: http://arxiv.org/abs/1907.10902.

68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python;12(85):2825–2830.

69. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. Association for Computing Machinery;. p. 785–794. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785.

6.3 Conclusion

Our analysis demonstrates that SHAP-based feature importance rankings can provide meaningful insights into transcriptional regulation. We observe that genes ranked highly by SHAP tend to correspond to direct regulatory targets, suggesting that this approach can be a useful hypothesis-generating tool for prioritizing candidate regulators. We validate the cooperative roles of SET1A and ZC3H4 at promoters and highlight a possible functional connection between ZC3H4 and INTS11, which may indicate crosstalk in transcription termination. While these findings align with prior knowledge, they also reveal patterns that warrant further experimental validation. This study underscores the potential of integrating predictive modelling with functional genomics to refine our understanding of transcriptional control.

Chapter 7

Discussion

This thesis applies machine learning and statistical modelling to high-throughput sequencing data to uncover the mechanistic roles of chromatin-associated proteins in transcriptional regulation. First, I challenged the global role of MeCP2 in alternative splicing, showing that its primary function lies in transcriptional regulation, with mCAC binding—not CA repeats—driving its activity. Second, I explored SALL4's role in transcriptional repression, revealing that its regulatory strength is directly proportional to the AT-content of target genes. Lastly, I leveraged SHAP-based explainable AI to infer regulatory interactions from deep learning models predicting Pol-II occupancy. Ranking genes by SHAP values predicted genes that overlapped significantly with those differentially regulated following rapid protein degradation, offering insights into regulators of transcription and their direct targets. This approach also identified a potential link between ZC3H4 and INTS11, suggesting possible crosstalk between the Restrictor and Integrator complexes in transcription termination. These findings present an interesting direction for further experimental validation.

In the following sections, I summarize the contributions of this thesis and discuss potential future directions, including the need for experimental validation of computational predictions, the integration of newer chromatin profiling techniques, and the broader implications of machine learning in deciphering transcriptional regulation.

7.1 Summary of contributions

7.1.1 Reevaluating MeCP2's Role in Splicing and Binding

This thesis re-evaluates the proposed role of MeCP2 in alternative splicing by applying quantitative modeling approaches to transcriptomic data from neuronal data sets (Section 4.1.2). By comparing differential gene expression and splicing changes following MeCP2 disruption, I demonstrate that MeCP2's primary function lies in transcriptional regulation rather than alternative splicing. My findings show that while MeCP2 binds to mCG and mCAC, it does not exhibit a global influence on alternative splicing. Instead, gene expression changes vastly outnumber differential splicing events, indicating that previous associations between MeCP2 and splicing may be indirect or context-dependent. Further, I examined MeCP2's binding specificity to CA repeats, testing the hypothesis that these sequences serve as primary MeCP2 targets

(Section 4.2.2). My analysis confirmed that mCAC binding is the key determinant of MeCP2 function, irrespective of whether it is found within CA repeats or elsewhere in the genome. This finding refines our understanding of MeCP2's role in neuronal transcriptional regulation and suggests that its influence is governed by DNA methylation state rather than sequence motifs alone.

7.1.2 AT-rich DNA as a regulatory signal

This thesis provides novel insights into SALL4's function in transcriptional regulation, revealing that its repressive strength is directly proportional to the AT-content of target genes (Section 5.2). To establish this relationship, analysis was performed in SALL4 DNA-binding mutants and SALL4 overexpression cell lines, allowing us to systematically assess whether AT-rich genes were disproportionately affected by SALL4 levels. These experiments confirmed that genes with higher AT-content were more susceptible to repression when SALL4 was overexpressed, while loss of SALL4 binding led to their upregulation, reinforcing the idea that SALL4-mediated repression is inherently linked to DNA sequence composition.

7.1.3 Inferring regulatory interactions in transcriptional regulation

To uncover regulatory interactions among chromatin-associated proteins, this thesis applies SHAP-based explainable AI to deep learning models predicting RNA Pol-II occupancy (Section 6.2). I demonstrated that SHAP importance not only predicts differential gene expression with high accuracy but also captures the magnitude of transcriptional changes, using models trained from data in wildtype unperturbed conditions. This highlights the potential of machine learning models to identify key regulatory factors without requiring direct perturbation experiments. A key discovery from this approach was the potential functional link between ZC3H4 and INTS11, suggesting crosstalk between the Restrictor and Integrator complexes in transcription termination. My findings align with independent studies identifying overlapping target genes (Mimoso et al. 2025), further supporting the idea that ZC3H4 and INTS11 may function in a coordinated manner to regulate transcriptional pausing and termination. Further experimental validation will be necessary to confirm the extent and mechanism of this interaction.

7.2 Outlook

7.2.1 Is MeCP2 multi-functional?

Since the publication of my manuscript questioning MeCP2's global role in alternative splicing (Section 4.1.2), several studies have presented evidence suggesting a functional link between MeCP2 and splicing regulation. For instance, MeCP2 has been reported as a subunit of the Rbfox/LASR complex, which influences alternative splicing (Jiang et al. 2021). In this study, alternative splicing appears to be dysregulated following MeCP2 depletion and Rbfox mutation, but the observed splicing changes are based on RNA-seq data from cultured cells rather than neuronal tissues. Additionally, the relationship between DNA methylation, MeCP2 levels, and splicing changes is not explicitly modelled to establish a mechanistic link. Furthermore, proteomic studies have revealed a marked GO enrichment of MeCP2-associated proteins involved in RNA splicing (Liu et al. 2024). Since this work suggests that MeCP2 physically interacts with Pol–II, it raises the possibility that its association with splicing factors is indirect, mediated through the transcription machinery. Beyond splicing, new insights into MeCP2 chromatin binding have also emerged (Ibid.). Cleavage Under Targets and Tagmentation (CUT&Tag) experiments indicate that MeCP2 is predominantly enriched at promoters, a surprising finding given that CGIs at promoters are typically unmethylated. This contrasts with the broad chromatin binding patterns observed in earlier ChIP-seq studies (Figure 2 in Section 4.2.2), instead suggesting sharper peaks at promoters and enhancers (Mishra et al. 2025). With the advent of newer techniques such as CUT&RUN and CUT&Tag, which show significant deviations from ChIP-seq, additional modelling and analysis are needed to determine whether previously established relationships between DNA methylation and MeCP2 binding remain consistent. In summary, a complete understanding of MeCP2 biology remains an ongoing challenge.

Another promising direction for future research is the development of inducible degron-based cell lines to facilitate rapid degradation of MeCP2. Such systems could provide precise temporal control over MeCP2 levels, helping to disentangle its immediate transcriptional effects from secondary adaptations. This new wave of MeCP2-focused research underscores that many mechanistic questions remain unanswered, and further investigation is required to fully understand its biological role. From a clinical perspective, efforts to restore MeCP2 function in RTT are gaining momentum. A proof-of-principle study demonstrated that re-expressing wild-type MeCP2 in the brain rescues defective phenotypes in MeCP2-null animals (Guy et al. 2007). Inspired by this finding, many laboratories are actively exploring therapeutic

strategies for RTT, and two studies have recently progressed to clinical trials (Palmieri et al. 2023). This growing body of research is crucial in the quest to establish a viable cure for RTT. Even with potential therapeutic advances, a deeper understanding of MeCP2's function will be key to ensuring long-term efficacy and uncovering its broader role in neurobiology.

7.2.2 SALL4's role in transcription

One open question is whether SALL4's recruitment of the NuRD co-repressor complex is linked to its preference for AT-rich DNA. Given that NuRD remodels chromatin to repress transcription, its association with SALL4 at AT-rich loci could provide insight into how DNA sequence composition influences chromatin accessibility and gene silencing. Further research using high-resolution chromatin profiling and inducible degradation approaches will be necessary to determine whether SALL4-NuRD interactions are context-dependent or specifically enriched at AT-rich sites. Understanding this relationship could shed light on how global DNA features shape transcriptional control and lineage commitment.

7.2.3 Enhancing SHAP analysis for transcriptional regulation

While this thesis highlights the power of SHAP-based explainable AI in uncovering regulatory interactions in transcription, several limitations remain. A key challenge is that current models rely primarily on chromatin-associated protein occupancy and histone modifications, without incorporating DNA sequence features. Given that regulatory elements such as CGIs and TFs motifs influence gene expression, future models should integrate sequence-based features alongside chromatin data. Doing so could help differentiate whether a protein's influence on transcription is due to direct binding at a sequence-specific site or indirect recruitment through chromatin interactions.

Additionally, while SHAP assigns feature importance, it does not inherently distinguish direct from indirect regulatory effects. Recent advancements in causal Shapley values (Heskes et al. 2020), which leverage *do*-calculus (Pearl 2012) to separate direct and indirect effects, provide a powerful framework for refining SHAP-based analyses by ensuring feature attributions remain robust even in the presence of feature dependencies. Applying causal SHAP techniques to transcriptional perturbation datasets particularly those using inducible degron systems could help refine which chromatin-associated proteins are primary drivers of transcriptional changes versus secondary responders. Another key limitation is that this SHAP analysis has primarily focused on promoters and gene bodies, leaving enhancer loci unexplored. Enhancers play a critical role in long-range transcriptional regulation and act as important

regulators of gene expression. Extending SHAP-based analysis to incorporate enhancer landscapes and chromatin loops would provide a more holistic view of transcriptional regulation. Moreover, extending SHAP analysis to predict transcriptional elongation and termination processes through nascent RNA sequencing would allow for a more complete view of how regulatory factors influence different stages of gene expression. By addressing these challenges, future research can expand the applicability of SHAP in transcriptional biology, bridging the gap between data-driven inference and mechanistic understanding.

Chapter A

Additional methods and source code

This appendix contains supplementary material for the papers included in this thesis, as well as source code for the bioinformatics workflows and models used throughout the research. For each paper, additional supplementary figures, tables and detailed extensive methods are provided to support the findings presented in the main text. The source code is made available to ensure reproducibility and transparency in the computational methods employed.

S1 Appendix. Detailed information about software, software versions and command line arguments used for analysis

Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing

Kashyap Chhatbar, Justyna Cholewa-Waclaw, Ruth Shah, Adrian Bird, Guido Sanguinetti

DNA methylation regions & ChIP-seq signal regions

Region specific mean methylation, methylation density and ChIP signal are calculated for the regions described in Fig A

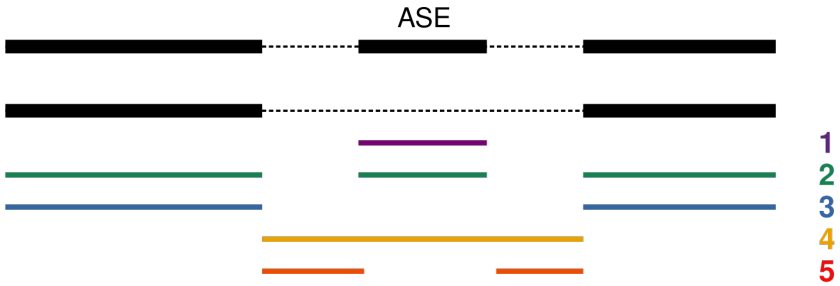


Fig A. Regions of gene for which DNA methylation metrics are calculated 1) ASE (alternative spliced exon) 2) combination of all exons containing the ASE 3) combination of all exons without the ASE 4) combination of all introns 5) combination of all introns without the ASE.

Genome assembly and annotation

Homo sapiens assembly (hg19) is used for the GSE125660 [9] and *Mus musculus* assmebly (mm10) is used for GSE128186 [17], GSE103214 [20], GSE48307 [23] and GSE85517 [14]. The genome annotation for alternative splicing events are made available as part of BRIE [22] and downloaded from `https://sourceforge.net/projects/brie-rna/files/annotation/`. HISAT2 [29] genome indexes for hg19 and mm10 are downloaded from `ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/data`.

Software and command-line arguments

After downloading the raw reads from Gene Expression Omnibus archive, we used HISAT2 [29] v2.0.2 for aligning the reads to the genome using the following command.

```
hisat2 -x {genome.index} -1 {input.one} -2 {input.two} --dta-cufflinks --known-splicesite-
    infile {splicesite.file} -k 2 --no-unal -p 4 | samtools view -h -b -@ 4 -o {output} -
```

Miso [27] v0.5.4 is used to quantify splicing ratios from the aligned files. The `indexed.assembly` is generated from genome annotation for alternative splicing events using the command `index_gff` which is part of Miso.

```
miso --run {indexed.assembly} {file.bam} --output-dir {outdir} --paired-end {mean.fragment.
    length} {stdev.fragment.length} --settings-filename miso_settings.txt --read-len {read.
    length}
```

Miso [27] v0.5.4 is used to estimate differential splicing between samples. Alignments from different replicates are pooled into a single BAM file and then used for comparison.

```
compare_miso --compare-samples {treatment} {control} {outdir}
```

BRIE [22] v0.2.0 is used to quantify splicing ratios from the aligned files.

```
brie -a {gtf} -s {file.bam} -f {brie.features} -o {outdir}
```

Cufflinks [30] v2.2.1 is used to quantify splicing ratios from the aligned files.

```
cufflinks -G {gtf} -o {outdir} {file.bam}
```

Differential gene expression is estimated using DESeq2 [32] v1.24.0 with genotype or age as the covariate in the `design` formula.

Generating Simulation Data

Sequencing reads were generated using Art `https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm` on reference RNA sequences from Sequins `https://www.sequinstandards.com` at different coverage depths.

```
art_illumina -na -i RNAsequins.v2.2.fa -l 100 -ss HS25 -f 25 -m 200 -s 10 -o fastq/sim1.
art_illumina -na -i RNAsequins.v2.2.fa -l 100 -ss HS25 -f 50 -m 200 -s 10 -o fastq/sim2.
art_illumina -na -i RNAsequins.v2.2.fa -l 100 -ss HS25 -f 100 -m 200 -s 10 -o fastq/sim3.
```

Reads were then aligned using HISAT2 [29] as described earlier and Splicing Ratios were quantified using Miso [27] and Cufflinks [30]. Splicing Ratios were then visualized as in S3 Fig D.

Regressing Sequence Features with Splicing Ratios

The modelling with Sequence Features and differences in DNA methylation patterns with differences in Splicing Ratios is executed using the same way as described below.

```
from scipy import stats
from sklearn.model_selection import cross_validate
from sklearn import linear_model
from sklearn import ensemble

import glob
import pandas as pd

miso_summaries = glob.glob("*.miso_summary")
features = pd.read_csv("factors.csv.gz", compression="gzip", index_col=0)

r2_scores = []

for summary in miso_summaries:
        sample = summary.split(".miso_summary")[0]
        psi = pd.read_csv(fraction, index_col=0, usecols=[0,1], sep="\t").replace(0,0.00001).
            replace(1, 0.99999).apply(stats.norm.ppf)
        y = psi["miso_posterior_mean"]
        X = features.loc[y.index]

        lm = linear_model.BayesianRidge(n_iter=1000, tol=0.00001, normalize=True,
            compute_score=True)
```

```
        lm_scores = cross_validate(lm, X, y, cv=5, scoring=('r2'), n_jobs=-1)
        for score in lm_scores["test_scores"]:
                r2_scores.append(["Linear Regression", "Miso", sample, score])

        rf = ensemble.RandomForestRegressor(n_estimators=100, random_state=0, n_jobs=-1)
        rf_scores = cross_validate(rf, X, y, cv=5, scoring=('r2'), n_jobs=-1)
        for score in rf_scores["test_scores"]:
                r2_scores.append(["Random Forest Regression", "Miso", sample, score])

brie_fractions = glob.glob("*.fractions.tsv")

for fraction in brie_fractions:
        sample = summary.split(".fractions.tsv")[0]
        psi = pd.read_csv(fraction, index_col=0, sep="\t", usecols=[0,5]).replace(0,0.00001).
            replace(1, 0.99999).apply(stats.norm.ppf)
        X = features.copy()
        X.index = ["{}.in".format(ii) for ii in X.index]
        y = psi.loc[X.index]["Psi"]

        lm = linear_model.BayesianRidge(n_iter=1000, tol=0.00001, normalize=True,
            compute_score=True)
        lm_scores = cross_validate(lm, X, y, cv=5, scoring=('r2'), n_jobs=-1)
        for score in lm_scores["test_scores"]:
                r2_scores.append(["Linear Regression", "BRIE", sample, score])

        rf = ensemble.RandomForestRegressor(n_estimators=100, random_state=0, n_jobs=-1)
        rf_scores = cross_validate(rf, X, y, cv=5, scoring=('r2'), n_jobs=-1)
        for score in rf_scores["test_scores"]:
                r2_scores.append(["Random Forest Regression", "BRIE", sample, score])
```

Scatter Plot

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import gaussian_kde

df = pd.read_csv("y_test_vs_y_pred.tsv", sep="\t", index_col=0)
fig, ax = plt.subplots()
y_te = df["y_test"]
y_pr = df["y_pred"]
xy = np.vstack([y_te, y_pr])
z = gaussian_kde(xy)(xy)
ax.scatter(y_te, y_pr, s=30, c=z, edgecolor='', cmap="Greys")
ax.set_xlim(-3, 3)
ax.set_ylim(-3, 3)
ax.set_xticks([-2, 0, 2])
ax.set_yticks([-2, 0, 2])
ax.plot(bx.get_xlim(), bx.get_ylim(), ls="--", c=".3")
```

Regression Analysis Overview

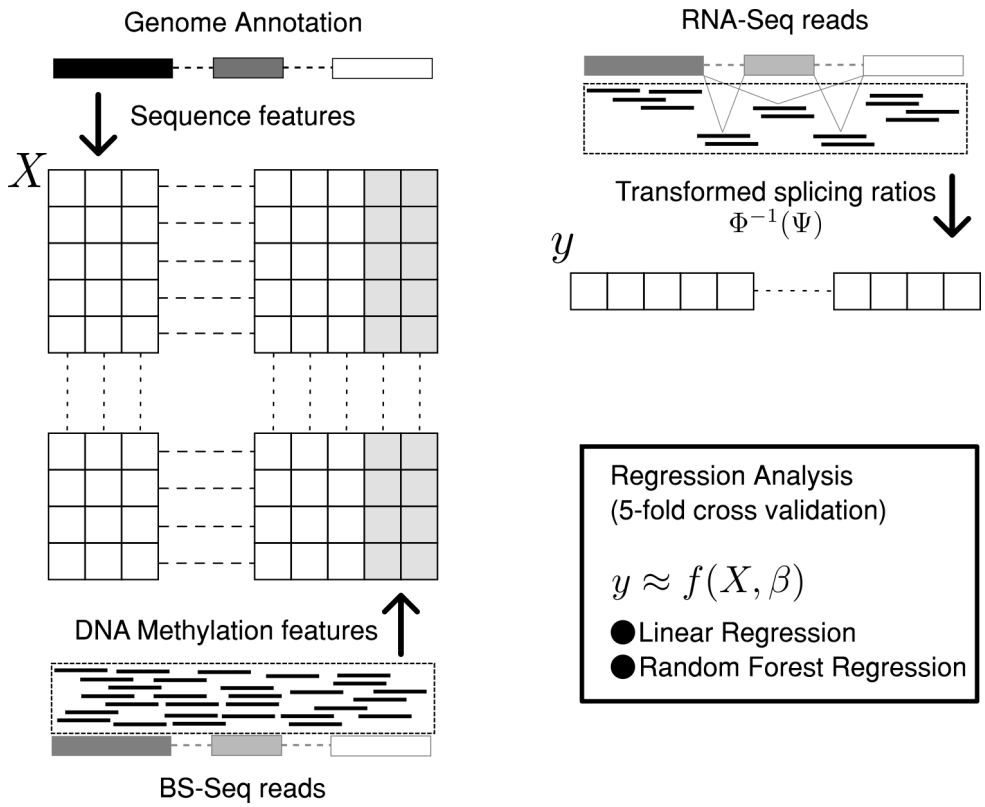


Fig B. Graphical representation of the overall regression analyses. Accumulation of sequence features from genome annotation, accumulation of DNA Methylation features from BS-Seq (Bisulfite Sequencing) reads and quantification of Splicing Ratios (Ψ) from RNA-Seq (RNA Sequencing) reads.

Molecular Cell, Volume *81*

Supplemental Information

SALL4 controls cell fate

in response to DNA base composition

Raphaël Pantier, Kashyap Chhatbar, Timo Quante, Konstantina Skourti-Stathaki, Justyna Cholewa-Waclaw, Grace Alston, Beatrice Alexander-Howden, Heng Yang Lee, Atlanta G. Cook, Cornelia G. Spruijt, Michiel Vermeulen, Jim Selfridge, and Adrian Bird

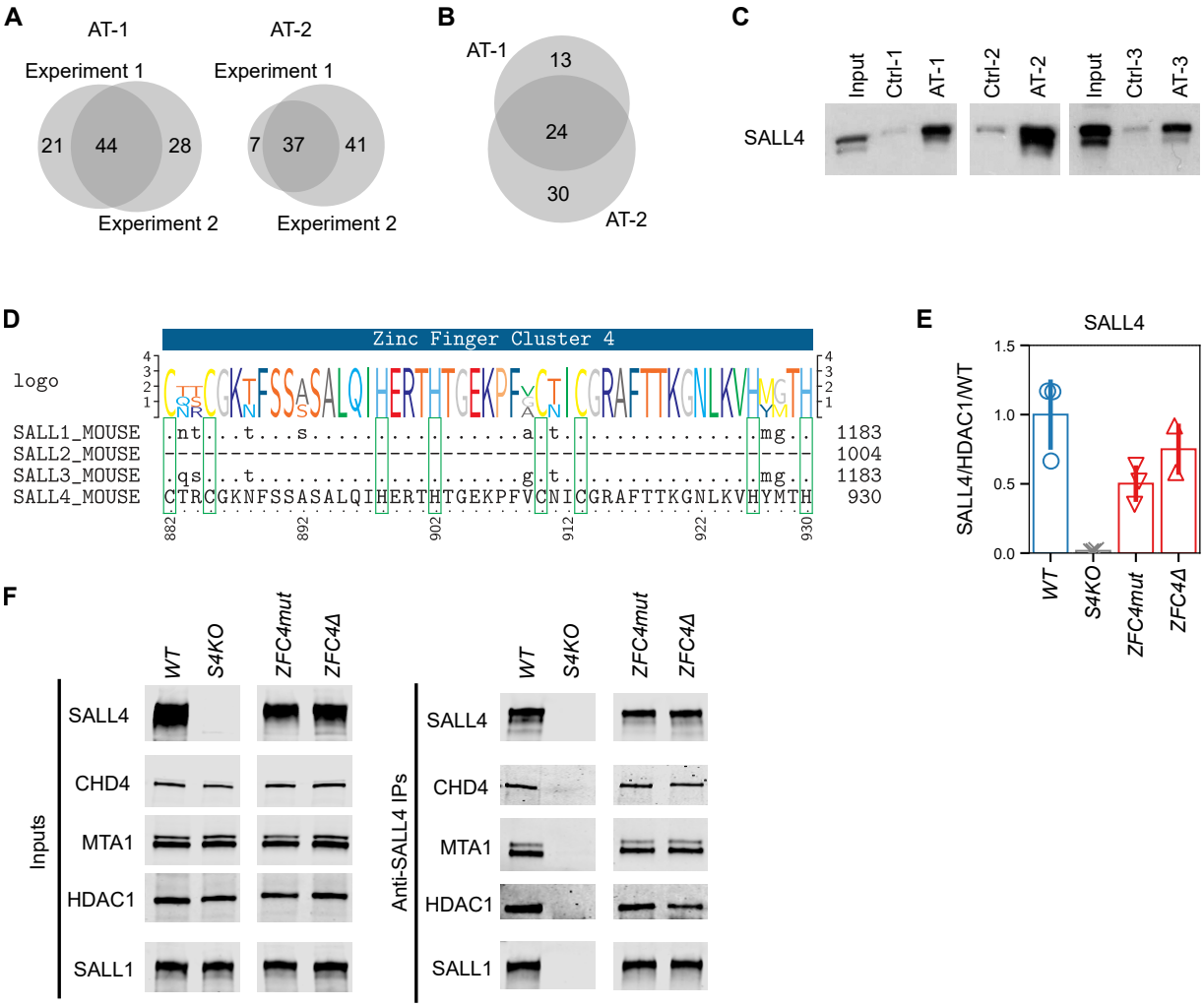


Figure S1: Identification of novel AT-binding proteins in embryonic stem cells by DNA pulldown-mass spectrometry (related to Figure 1)

A, B. Venn diagrams showing the overlap between proteins identified by DNA pulldown-mass spectrometry in independent replicate experiments (A), or using unrelated AT-rich DNA probes (B). **C.** DNA pulldown with AT-rich (AT-1, AT-2, AT-3) or control (Ctrl-1, Ctrl-2, Ctrl-3) probes followed by Western blot analysis for SALL4 using *WT* ESC protein extracts. **D.** Protein alignment and consensus sequence of C2H2 zinc-finger cluster 4 (ZFC4) in the mouse SALL protein family. ZFC4 is absent in SALL2. **E.** Western blot quantification of SALL4 expression levels in *S4KO* and *ZFC4mut/*Δ ESCs, normalised to HDAC1 expression and relative to *WT* ESC levels. Data points indicate independent replicate experiments and error bars standard deviation. **F.** SALL4 co-immunoprecipitation with SALL1 and NuRD components in *WT*, *S4KO* (negative control) and *ZFC4mut/*Δ ESCs. For both inputs and anti-SALL4 IPs, all four lanes are part of the same Western blot membrane and images were processed in an identical manner.

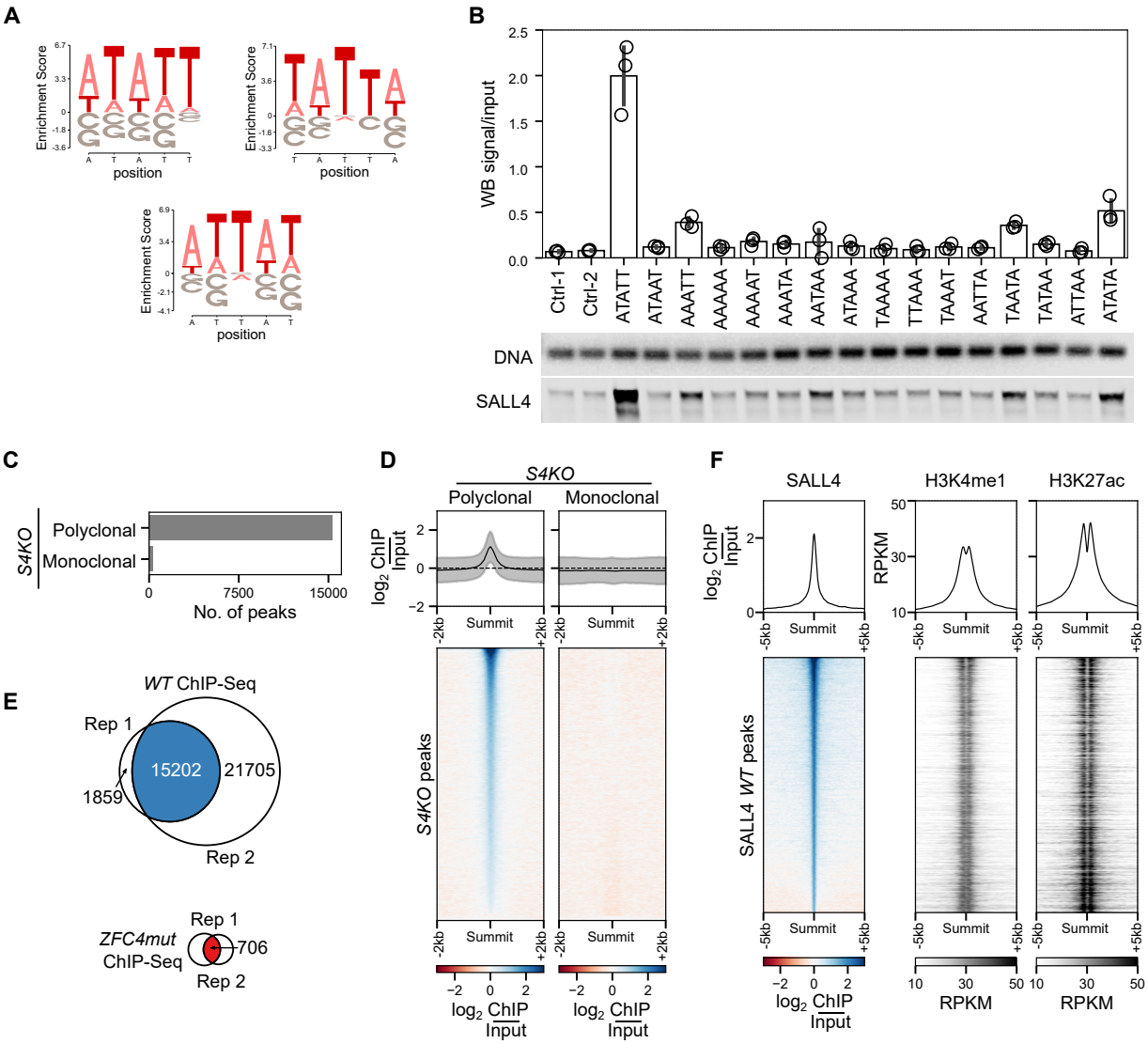


Figure S2: Characterisation of SALL4 C2H2 zinc-finger cluster 4 (ZFC4) DNA binding *in vitro* and *in vivo* (related to Figure 2)

A. Motif logos generated from the three most enriched k-mers (n=5) after 6 cycles of HT-SELEX with SALL4 ZFC4. **B.** DNA pulldown with AT-rich probes containing all possible combinations of AT 5 mers or control probes with disrupted AT-runs (Ctrl-) followed by Western blot analysis for SALL4. Amounts of DNA probes were assessed by agarose gel analysis and SALL4 enrichment was normalised to input. Data points indicate independent replicate experiments and error bars standard deviation. **C.** Detection of non-specific SALL4 ChIP-seq peaks in *Sall4* knockout ESCs (negative control) using either a monoclonal or a polyclonal anti-SALL4 antibody. **D.** Profile plot and heatmap showing SALL4 ChIP-seq signal in *Sall4* knockout ESCs at non-specific sites (see panel B) using either a monoclonal or a polyclonal anti-SALL4 antibody. **E.** Venn diagrams showing the overlap of SALL4 ChIP-seq peaks between independent replicate experiments using an anti-SALL4 monoclonal antibody in *WT* (blue) and *ZFC4mut* (red) ESC lines. **F.** Profile plots and heatmaps showing SALL4, H3K4me1 and H3K27ac ChIP-seq signal at SALL4 *WT* ChIP-seq peaks in *WT* ESCs.

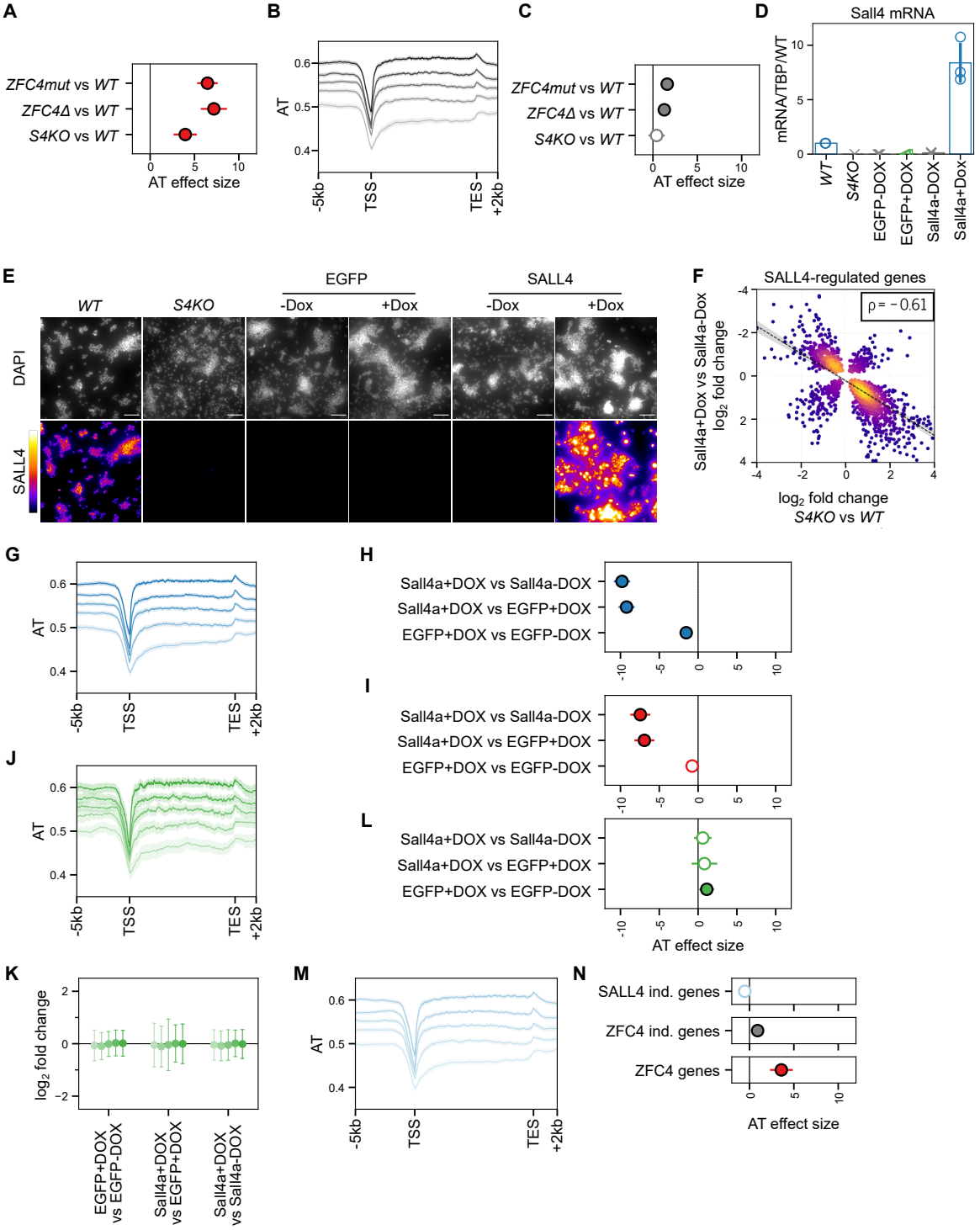


Figure S3: SALL4-mediated transcriptional regulation in relation to DNA base composition (related to Figure 3)

(legend continued on next page)

A. Statistical analysis of AT-dependent gene expression changes (coefficient estimates with 99% confidence intervals) observed with ZFC4-regulated genes (see Figure 3A). Significance is attributed by the F-test. Empty circles represent non-significant model fits (>0.01 FDR) and filled circles represent a significant fit to the model. **B.** Profile plot showing the density of A/T nucleotides around the transcription unit of ZFC4-independent genes (see Figure 3A) divided into five equal categories according to AT-content. **C.** Statistical analysis of AT-dependent gene expression changes observed with ZFC4-independent genes, as described in panel A. **D.** RT-qPCR analysis following 48h doxycycline induction in the indicated ESC lines (see Figure 3E), or in *WT* and *S4KO* control ESCs. Sall4 mRNA expression was normalised to TBP and expressed relative to *WT*. Data points indicate independent replicate experiments and error bars standard deviation. **E.** SALL4 immunofluorescence following 48h doxycycline induction in the indicated ESC lines (see Figure 3E), or in *WT* and *S4KO* control ESCs. DNA was stained with DAPI. Scale bars: 100µm. **F.** Scatter plot showing the relative expression of genes deregulated both in *S4KO* ESCs and following SALL4 re-expression. **G.** Profile plot showing the density of A/T nucleotides around the transcription unit of Sall4-responsive genes (see Figure 3F) divided into five equal categories according to AT-content. **H, I.** Statistical analysis of AT-dependent gene expression changes observed with Sall4-responsive (H) and ZFC4-regulated (I) genes, as described in panel A. **J.** Profile plot showing the density of A/T nucleotides around the transcription unit of EGFP-responsive genes (see Figure 3F) divided into five equal categories according to AT-content. **K.** Correlation between EGFP-induced gene expression changes and DNA base composition. EGFP-responsive genes were divided into five equal categories depending on their AT-content, and their relative expression levels were analysed in the indicated ESC lines. **L.** Statistical analysis of AT-dependent gene expression changes observed with EGFP-responsive genes, as described in panel A. **M.** Profile plot showing the density of A/T nucleotides around the transcription unit of SALL4-independent genes changing during early ESC differentiation (see Figure 3J) divided into five equal categories according to AT-content. **N.** Statistical analysis of AT-dependent gene expression changes observed with SALL4-independent genes (light blue), SALL4-dependent genes controlled by ZFC4 (red) and SALL4-dependent genes not controlled by ZFC4 (grey) during early differentiation of *WT* cells (day 0 *vs* day 2), as described in panel A.

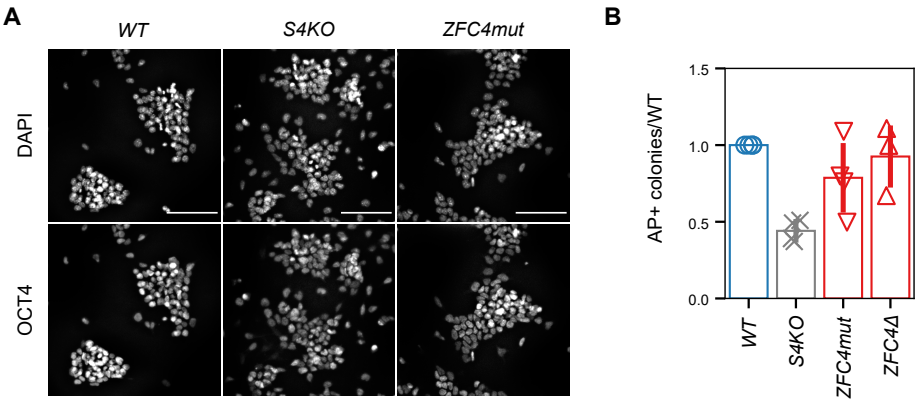


Figure S4: Phenotypic effects of SALL4 ZFC4 mutation on neuronal differentiation (related to Figure 4)

A. OCT4 immunofluorescence in *WT*, *S4KO* and *ZFC4mut* ESCs. DNA was stained with DAPI. Scale bars: 100µm. **B.** Self-renewal assay in *WT*, *S4KO* and *ZFC4mut/*Δ ESCs. Alkaline phosphatase (AP)-positive colonies were counted and normalised to *WT*. Data points indicate independent replicate experiments and error bars standard deviation.

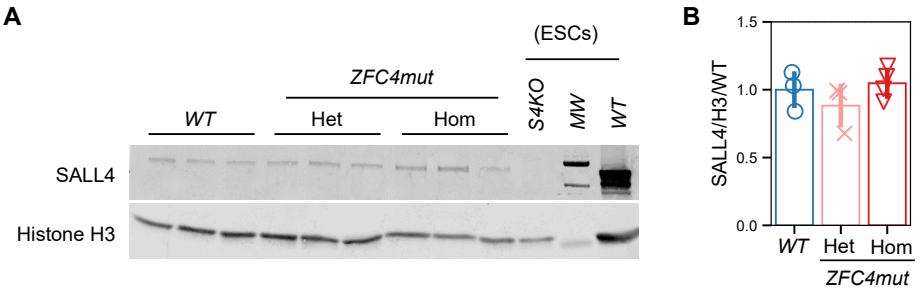


Figure S5: Mutation of SALL4 ZFC4 causes embryonic lethality (related to Figure 5)

A. Western blot analysis of SALL4 in *WT*, *ZFC4mut* heterozygote (*Het*) and homozygote (*Hom*) embryos at E10.5. *WT* and *S4KO* ESC protein extracts were used as controls. **B.** Western blot quantification of SALL4 expression levels in *ZFC4mut* embryos (as presented in panel A), normalised to Histone H3 expression and relative to *WT*. Data points indicate independent embryos and error bars standard deviation.

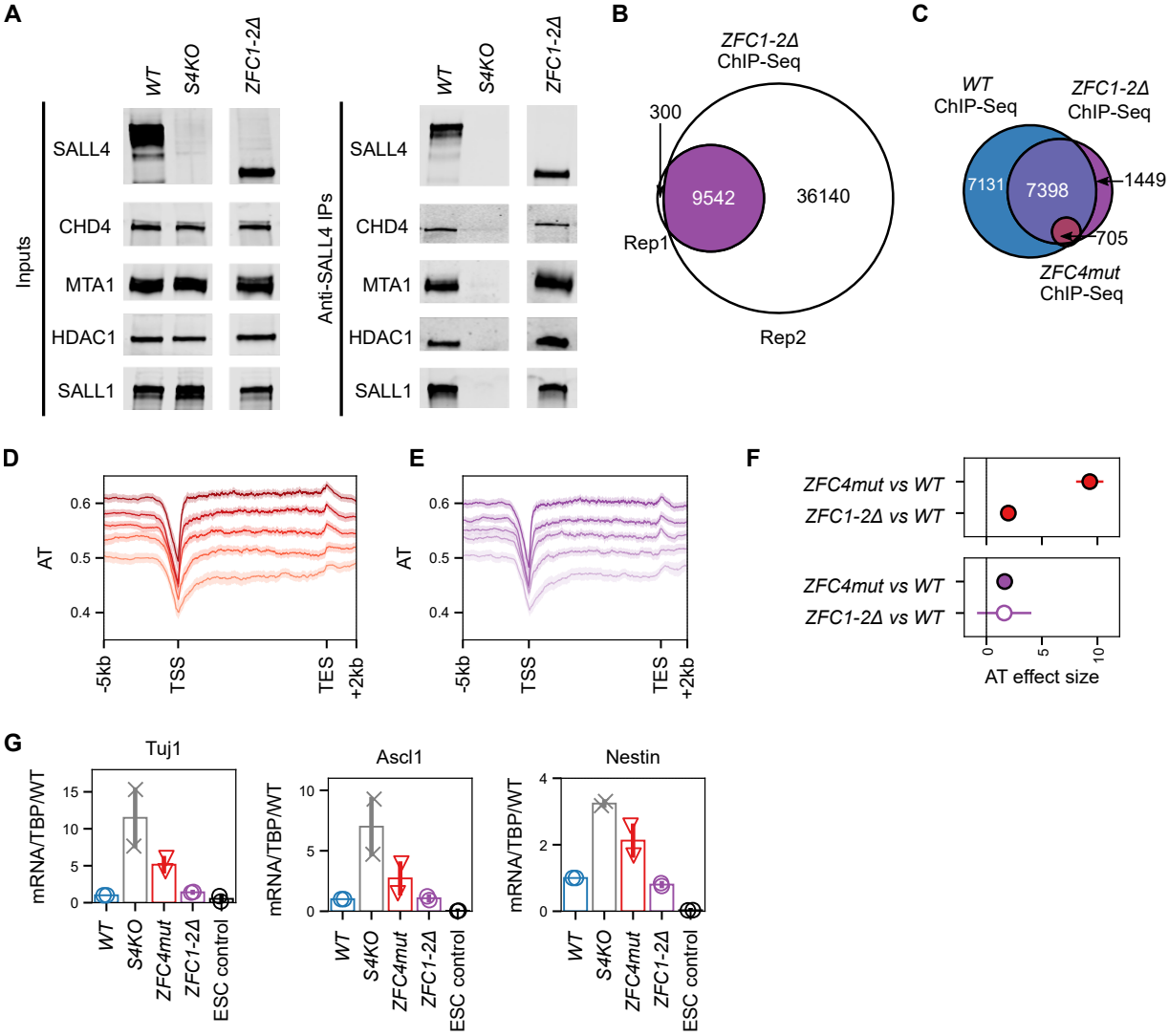


Figure S6: Effects of SALL4 ZFC1-2 deletion in ESCs on chromatin binding, gene expression and differentiation (related to Figure 6)

A. SALL4 co-immunoprecipitation with SALL1 and NuRD components in *WT*, *S4KO* (negative control) and *ZFC1-2Δ* ESCs. For both inputs and anti-SALL4 IPs, all three lanes are part of the same Western blot membrane and images were processed in an identical manner. **B.** Venn diagram showing the overlap of SALL4 ChIP-seq peaks between independent replicate experiments in *ZFC1-2Δ* ESCs. **C.** Venn diagram showing the overlap of SALL4 ChIP-seq peaks between *WT*, *ZFC1-2Δ* and *ZFC4mut* ESCs. **D, E.** Profile plot showing the density of A/T nucleotides around the transcription unit of ZFC4-regulated (D) and ZFC1/2-regulated (E) genes (see Figure 6E) divided into five equal categories according to AT-content. **F.** Statistical analysis of AT-dependent gene expression changes (coefficient estimates with 99% confidence intervals) observed with ZFC4-regulated (red) and ZFC1/2-regulated (purple) genes (see Figure 6E). Significance is attributed by F-test. Empty circles represent non-significant model fits (>0.01 FDR) and filled circles represent significant model fit. **G.** RT-qPCR analysis of the neuronal markers Tuj1, Ascl1 and Nestin in the indicated cell lines following differentiation for 5 days in N2B27 medium. Transcripts levels were normalised to TBP and expressed relative to *WT*. Data points indicate independent replicate experiments and error bars standard deviation.

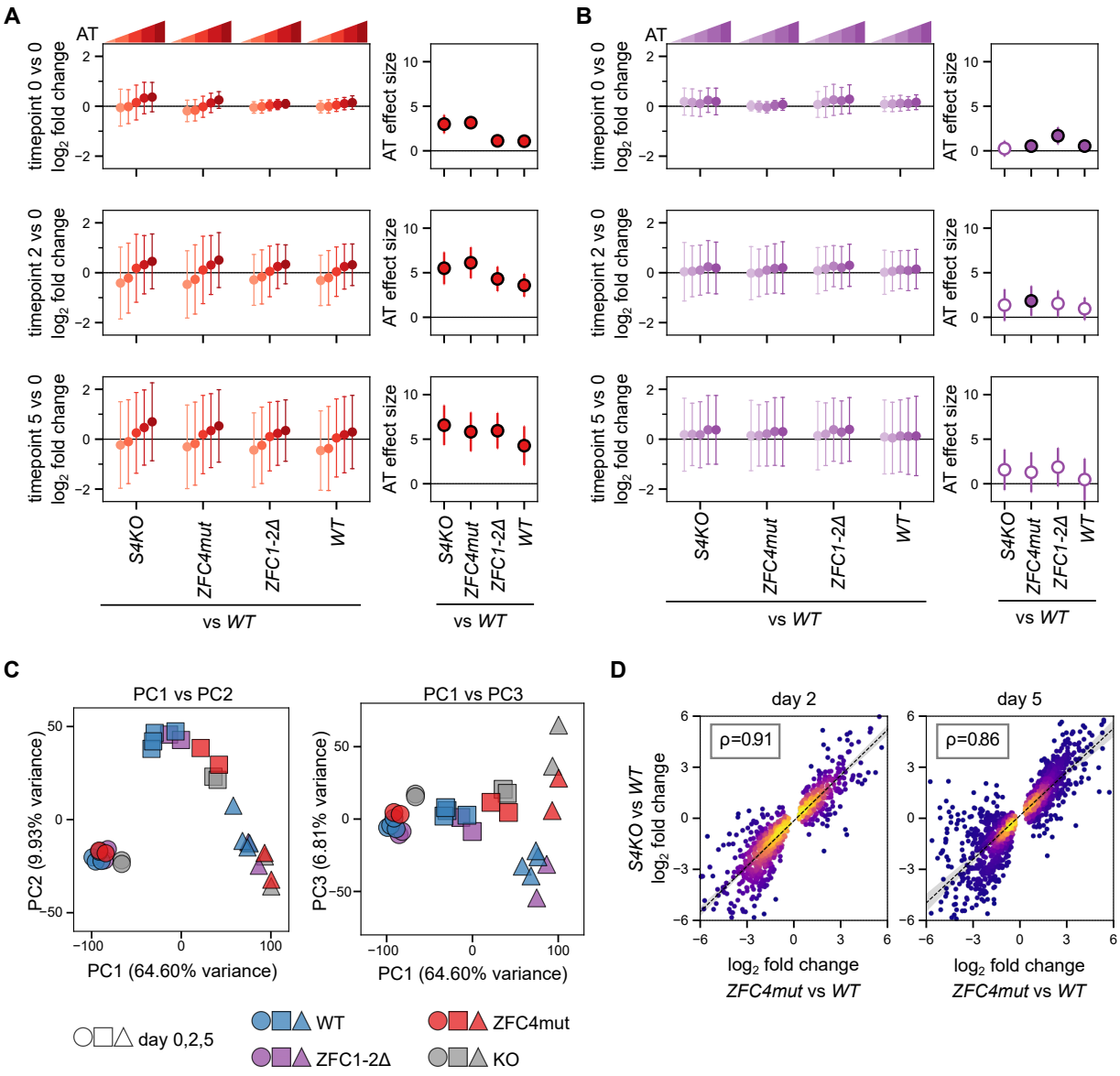


Figure S7: Transcriptional effects of SALL4 zinc finger cluster mutations during neuronal differentiation (related to Figure 7)

A. Correlation between gene expression changes and DNA base composition observed with ZFC4-regulated genes at day 0 (top panel), day 2 (middle panel) and day 5 (bottom panel) of differentiation. ZFC4-regulated genes (see Figure 6E) were divided into five equal categories according to their AT-content. Left panel: relative expression levels (log2 fold-change *vs* day 0 in *WT* cells) in *WT* and *Sall4* mutant cells. Right panel: Coefficient estimates (with 99% confidence intervals) describing the AT effect size. **B.** Correlation between gene expression changes and DNA base composition observed with ZFC1/2-regulated genes during differentiation, as described in panel A. **C.** PCA analysis of RNA-seq samples from *WT* and *Sall4* mutant cell lines at day 0, 2 and 5 of differentiation. **D.** Scatter plot showing the relative expression levels of genes deregulated in differentiating *ZFC4mut* cells (see Figure 7B, red bars) correlating with their expression in *S4KO* cells at day 2 and 5 of differentiation.

Methods S1. Bioinformatics analysis - command line arguments (related to STAR Methods. Quantification and Statistical Analysis)

Command line arguments for counting k-mers

k-mer abundance was calculated using the following commands

```
jellyfish count -m 5 -C -t 4 -s 100M -o 5.jf <(zcat sample.fq.gz)

jellyfish dump 5.jf > 5_counts.fa
```

Fraction of reads containing k-mers was calculated after executing `calculate_fraction.py` and `calculate_score.py` scripts on counts obtained using the above steps. Analysis pipeline for executing scripts is included in deposited Mendeley data (DOI: 10.17632/rwzttj9pn2.1).

Command-line arguments for ChIP-seq anlaysis

```
trimmomatic SE -threads 16 -summary R1.trimmomatic.log R1.fq R1.trimmed.fq
    ↪ ILLUMINACLIP:adapters/TruSeq-SE-combined.fa:2:30:10 LEADING:3 TRAILING:3
    ↪ SLIDINGWINDOW:4:20 MINLEN:36

bwa mem -t 6 -M mm10 R1.trimmed.fq | samtools view -bT mm10.fa -q 1 -F 4 -F 256 >
    ↪ R1.unsorted.bam

samtools sort -o R1.sorted.bam R1.unsorted.bam

samtools index R1.sorted.bam

picard MarkDuplicates I=R1.sorted.bam O=R1.dedup.sorted.bam ASSUME_SORTED=true
    ↪ REMOVE_DUPLICATES=true METRICS_FILE=R1.dedup.metrics
    ↪ VALIDATION_STRINGENCY=LENIENT PROGRAM_RECORD_ID='null'

samtools index R1.dedup.sorted.bam

computeGCBias -b R1.dedup.sorted.bam --effectiveGenomeSize 2494787188 -g mm10.2bit
    ↪ -bl blacklist.bed -p 20 -l 240 -o R1.dedup.gcbias.freq --biasPlot
    ↪ R1.dedup.gcbias.png

correctGCBias -b R1.dedup.sorted.bam --effectiveGenomeSize 2494787188 -g mm10.2bit
    ↪ -p 20 -freq R1.dedup.gcbias.freq -o R1.dedup.sorted.gc.corrected.bam

samtools index R1.dedup.sorted.gc.corrected.bam

macs2 callpeak -t R1.chip.dedup.sorted.gc.corrected.bam -c
    ↪ R1.control.dedup.sorted.gc.corrected.bam -f BAM -g 2494787188 --outdir macs
    ↪ -n R1.chip

bamCompare -b1 R1.chip.dedup.sorted.gc.corrected.bam -b2
    ↪ R1.input.dedup.sorted.gc.corrected.bam --scaleFactorsMethod None
    ↪ --effectiveGenomeSize 2494787188 -p 10 --operation log2 --normalizeUsing RPKM
    ↪ -bl blacklist.bed -o R1.chip.input.log2.bw
```

```
computeMatrix reference-point -a 2000 -b 2000 --referencePoint center --smartLabels
    ↪ -R peaks.bed -S R1.chip.input.log2.bw -bs 5 -p 48

meme-chip -o R1.chip.meme -neg random.ATAC.fasta -order 2 -meme-p 12 -meme-nmotifs
    ↪ 40 -psp-gen R1.chip.peaks.fasta
```

Command-line arguments for RNA-seq analysis

```
sailfish quant -l IU -i gencode.M23.index -1 R1.1.fq -2 R1.2.fq --biasCorrect -g
    ↪ gencode.M23.genes --numBootstraps 20 -o outdir -p 12

bedtools makewindows -g GRCm38.p6.fa.fai -w 1000 -i srcwinnum > GRCm38.p6.1kb.bed

bedtools nuc -fi GRCm38.p6.fa -bed GRCm38.p6.1kb.bed > GRCm38.p6.1kb.nuc

computeMatrix scale-regions -m 10000 -a 2000 -b 5000 -R gencode.M23.genes.bed -S
    ↪ GRCm38.p6.AT.bw -out gencode.M23.genes.AT.matrix.gz
```

R Script for differential gene expression of *Sall4* mutants

```
library(BiocParallel)
library(DESeq2)
register(MulticoreParam(4))

deseq_function <- function(counts_file, design_file, threshold, out_prefix){
  counts = read.csv(counts_file, sep="\t", header = TRUE,
                    row.names = 1, check.names = FALSE)
  design = read.csv(design_file, header=TRUE, sep=",", row.names=1)

  dds <- DESeqDataSetFromMatrix(countData = counts,
                                colData = design,
                                design = ~ condition)

  dds <- dds[rowSums(counts(dds)) > threshold,]

  # Performing DESeq2 analysis
  dds <- DESeq(dds, parallel=TRUE)
  saveRDS(dds, file=paste(out_prefix, "dds.rds", collapse="", sep=""))
  rld <- rlog(dds)

  ko_vs_wt <- results(dds, c("condition", "KO", "WT"), independentFiltering = TRUE)
  write.table(as.data.frame(ko_vs_wt),
      file=paste( out_prefix, "ko_vs_wt.tsv", collapse = "", sep=""),
      quote=F, col.names=NA, sep="\t")
  print(paste(c(counts_file, "finished")))
}
```

R Script for analysing genotype-specific differences over time during stem cell differentiation

```
library(BiocParallel)
library(DESeq2)
register(MulticoreParam(4))

find_hull <- function(df) df[chull(df$PC1, df$PC2), ]

deseq_function <- function(counts_file, design_file, out_prefix){
  counts = read.csv(counts_file, sep="\t", header = TRUE,
                    row.names = 1, check.names = FALSE)
  design = read.csv(design_file, header=TRUE, sep="\t", row.names=1)
  design$name <- relevel(design$name, "WT")
  design$timepoint <- as.factor(design$timepoint)

  dds <- DESeqDataSetFromMatrix(countData = counts,
                                colData = design,
                                design = ~ name + timepoint + name:timepoint)

  # Performing DESeq2 analysis
  dds <- DESeq(dds, parallel=TRUE)
  saveRDS(dds, file=paste(out_prefix, "dds.rds", collapse="", sep=""))
  rld <- rlog(dds)

  ddsTC <- DESeq(dds, test="LRT", reduced = ~ name + timepoint)
  resTC <- results(ddsTC)

  write.table(assay(rld), file=paste(c(out_prefix, "rld.tsv"), collapse="", sep=""),
      ↪ sep="\t")
  write.table(as.data.frame(resTC), file=paste(c(out_prefix, "fc.tsv"), collapse="",
      ↪ sep=""), sep="\t")

  print(paste(c(counts_file, "finished")))
}
```

Linear Regression Model

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats import multitest

# Fitting OLS linear regression model to data
df = pd.read_csv("fold_change_AT.tsv", sep="\t")
X = sm.add_constant(df[["AT mean"]])
y = df["log2FoldChange"].values
model = sm.OLS(y, X).fit()
at_hi_conf, at_low_conf = tuple(model.conf_int(0.01).loc["AT mean"].T.values)
at_mean = model.params.loc["AT mean"]
r_squared = model.rsquared
f_pvalue = model.f_pvalue

# Adjusting Type I errors
_, combined_df["FDR"], _, _ = multitest.multipletests(combined_df["f_pvalue"].values,
    ↪ alpha=0.01, method="fdr_bh")
```

Supplementary methods

Modelling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies

Kashyap Chhatbar[1,2], Adrian Bird[2], Guido Sanguinetti[3]

1 School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
2 Institute of Cell Biology, University of Edinburgh, Edinburgh, United Kingdom
3 International School for Advanced Studies (SISSA), Trieste, Italy

* kashyap.c@ed.ac.uk

Sequencing data preprocessing

As described in Materials and Methods: ChIP-seq data analysis, the raw sequencing data for ChIP-seq experiments were downloaded from the Gene Expression Omnibus (GEO) and processed using a custom snakemake workflow (Supplementary Methods Fig 1A). The workflow includes the following steps:

1. Downloading raw sequencing data from GEO.
2. Aligning reads to the reference genome using Bowtie2 [1].
3. Removing reads mapping to ENCODE blacklisted regions to improve data quality.
4. Calling peaks and generating signal tracks using MACS2 [2].
5. Generating treatment and control pileup tracks.
6. Calculating fold enrichment tracks using the `bdgcmp` command.
7. Quantifying the signal using bigWigAverageOverBed [3] to calculate the mean signal across defined genomic regions, such as promoters and gene bodies.

For the Materials and Methods: RNA-seq and TT-seq data analysis, the raw sequencing data for RNA-seq and TT-seq experiments were also downloaded from GEO and processed using a custom snakemake workflow (Supplementary Methods Fig 1B). The workflow includes the following steps:

1. Downloading raw sequencing data from GEO.
2. Aligning reads to the reference genome using Bowtie2 [1].
3. Identifying and removing PCR duplicates using SAMtools markup [4] to eliminate artifacts due to library preparation.
4. Quantifying gene-level read counts using featureCounts [5] from the Subread package.
5. Performing differential gene expression analysis using DESeq2 [6] in R, applying default parameters to identify significantly differentially expressed genes.

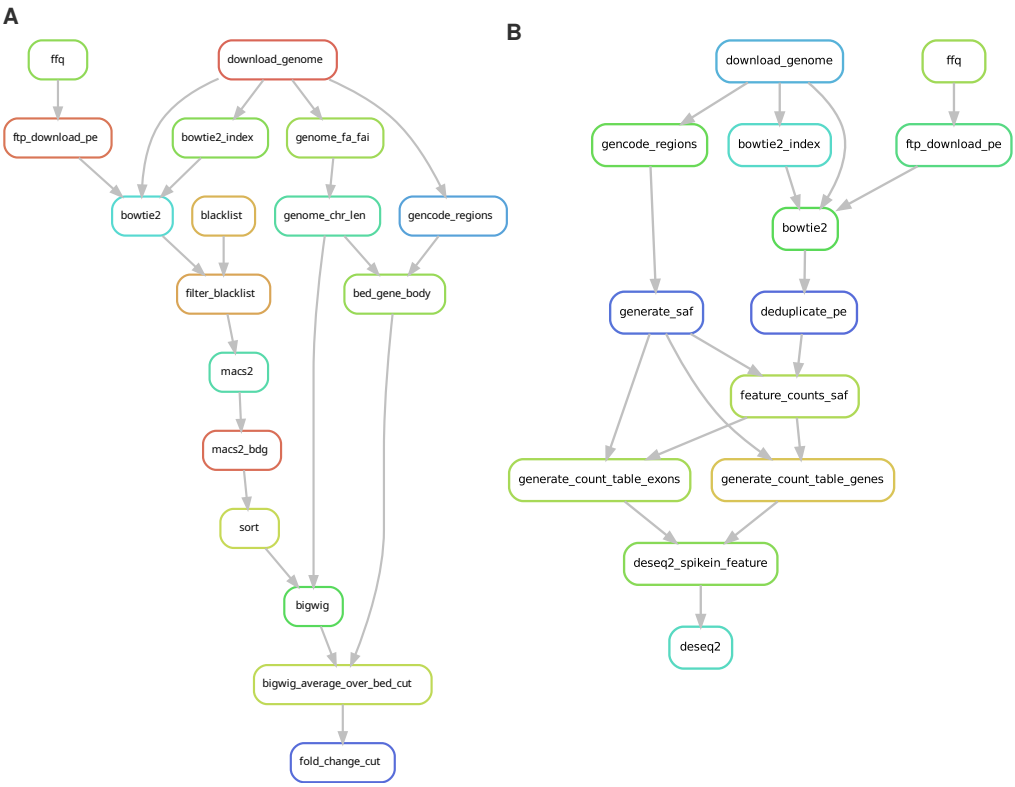
The DAGs (Directed Acyclic Graphs) of the snakemake workflows are shown in Supplementary Methods Fig 1. The complete workflow is publicly available on GitHub at https://github.com/kashyapchhatbar/SHAP-analysis.

Hyperparameter optimisation

The hyperparameter optimisation for the MLPRegressor and XGBoostRegressor models was performed using the Optuna library [7] as described in Materials and Methods: Data preparation and modelling. The following tables summarise the hyperparameters of the models trained on Hughes et al. [8], Dobrinić et al. [9] and Wang et al. [10] data.

Model training evaluation

The training process was performed again involving fitting the models to the training data and evaluating their performance using the best hyperparameters found during the optimisation. Supplementary Methods Fig 2 illustrates the training and validation MSE (mean squared error) losses and R^2 scores at each epoch for the PyTorch MLP models.



Supplementary Methods Fig 1. DAG of the snakemake workflow for ChIP-seq data analysis.

Supplementary Methods Table 1. Hyperparameters of the Scikit-learn MLPRegressor model trained on Hughes et al. [8] data.

l2_regularisation	hidden_layer_sizes_1	hidden_layer_sizes_2	activation
7.5167e-03	64	57	relu
1.3269e-05	62	56	relu
1.3797e-04	54	58	relu
9.6820e-05	45	61	relu
2.4365e-05	64	61	relu

Supplementary Methods Table 2. Hyperparameters of the PyTorch MLP model trained on Hughes et al. [8] data.

learning_rate_init	weight_decay	hidden_layer_sizes_1	hidden_layer_sizes_2
2.6827e-03	2.6086e-04	53	37
1.7020e-03	2.5392e-04	58	59
2.6403e-03	1.0215e-04	48	40
8.1475e-04	1.0007e-04	61	61
3.8772e-03	2.3777e-04	31	37

Feature correlation

To gauge how much shared signal exists among our chromatin features and whether this threatens interpretability, we calculated pair-wise Pearson correlations for every dataset used in modelling. The resulting matrices were rendered as clustered heatmaps to highlight groups of features that move together. For Hughes et al. [8] minimal three-feature

Supplementary Methods Table 3. Hyperparameters of the XGBoostRegressor model trained on Hughes et al. [8] data.

learning_rate_init	tree_method	max_depth
9.1785e-02	exact	6
7.9558e-02	exact	6
5.4488e-02	exact	6
6.5311e-02	exact	6
7.4898e-02	exact	7

Supplementary Methods Table 4. Hyperparameters of the Scikit-learn MLPRegressor model trained on Dobrinić et al. [9] data.

l2_regularisation	hidden_layer_sizes_1	hidden_layer_sizes_2	activation
4.5145e-05	61	64	relu
1.2066e-03	56	55	relu
1.4178e-04	62	64	relu
3.8390e-04	47	63	relu
1.2337e-03	60	58	relu

Supplementary Methods Table 5. Hyperparameters of the PyTorch MLP model trained on Dobrinić et al. [9] data.

learning_rate_init	weight_decay	hidden_layer_sizes_1	hidden_layer_sizes_2
6.9995e-04	2.1842e-04	61	51
2.5050e-03	4.3886e-04	56	23
2.2141e-03	1.4781e-04	60	27
4.0268e-03	2.5540e-04	46	62
6.9362e-03	3.9068e-04	42	51

Supplementary Methods Table 6. Hyperparameters of the XGBoostRegressor model trained on Dobrinić et al. [9] data.

learning_rate_init	tree_method	max_depth
7.4072e-02	hist	7
9.2142e-02	hist	6
6.2993e-02	hist	8
6.6056e-02	hist	7
6.4559e-02	hist	8

Supplementary Methods Table 7. Hyperparameters of the Scikit-learn MLPRegressor model trained on Wang et al. [10] data.

l2_regularisation	hidden_layer_sizes_1	hidden_layer_sizes_2	activation
1.9380e-02	60	61	relu
3.7036e-02	57	32	relu
2.7613e-02	48	59	relu
2.7805e-02	57	59	relu
1.6344e-02	63	56	relu

Supplementary Methods Table 8. Hyperparameters of the PyTorch MLP model trained on Wang et al. [10] data.

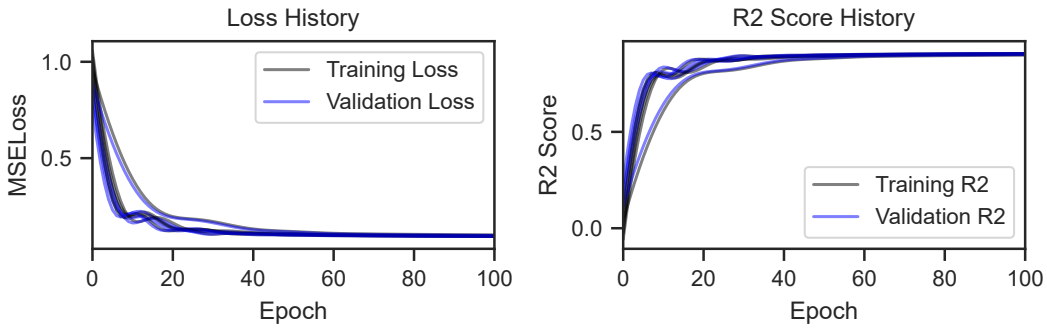
learning_rate_init	weight_decay	hidden_layer_sizes_1	hidden_layer_sizes_2
2.0964e-03	2.7503e-04	64	20
5.0577e-04	4.7558e-04	56	33
4.0241e-03	2.2214e-04	52	63
1.5923e-03	1.4617e-04	59	29
3.9265e-03	4.4898e-04	62	27

Supplementary Methods Table 9. Hyperparameters of the XGBoostRegressor model trained on Wang et al. [10] data.

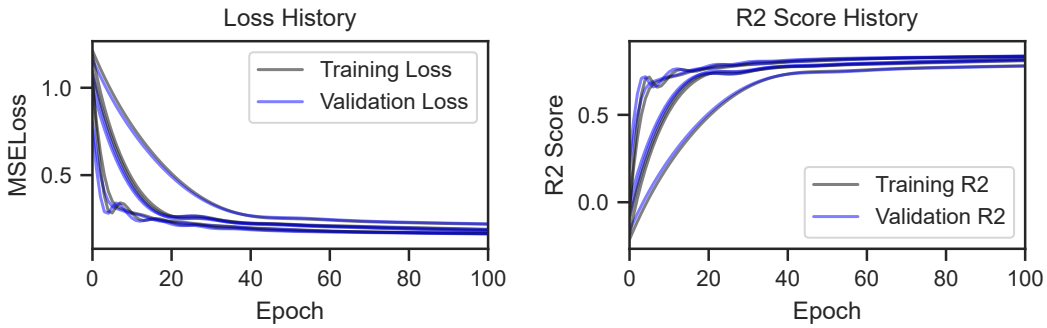
learning_rate_init	tree_method	max_depth
5.0245e-02	exact	8
6.5811e-02	hist	7
6.7234e-02	exact	7
6.7051e-02	hist	7
5.7773e-02	exact	9

model (SET1A, ZC3H4, H3K4me3), the results shows substantial positive correlations. Promoter (p) signals cluster separately than gene body (gb) signals for SET1A and ZC3H4, supporting our decision to model the two regions separately (Supplementary Methods Fig 3A). In Dobrinic et al. [9], promoter and gene body signals form separate clusters with an isolated H3K27me3 gene body signal (Supplementary Methods Fig 3B). We see high multi-colinearity in Wang et al. [10] data and separate clusters of promoter and gene body (Supplementary Methods Fig 3C). Mean absolute SHAP values across all splits are highly consistent, and the rank order of feature importance was identical across DeepSHAP, KernelSHAP and TreeSHAP. Because TreeSHAP leverages the tree structure, it is immune to the independence assumption that underlies KernelSHAP, providing an orthogonal check on the explanations. Together, these tests show that SHAP still identifies biologically meaningful drivers even in the presence of moderate to high feature correlations.

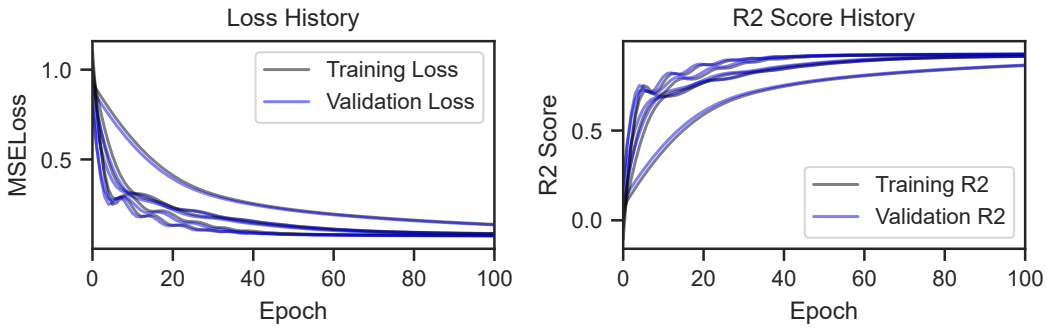
A Metrics for PyTorch MLP model trained on Hughes et al. data set



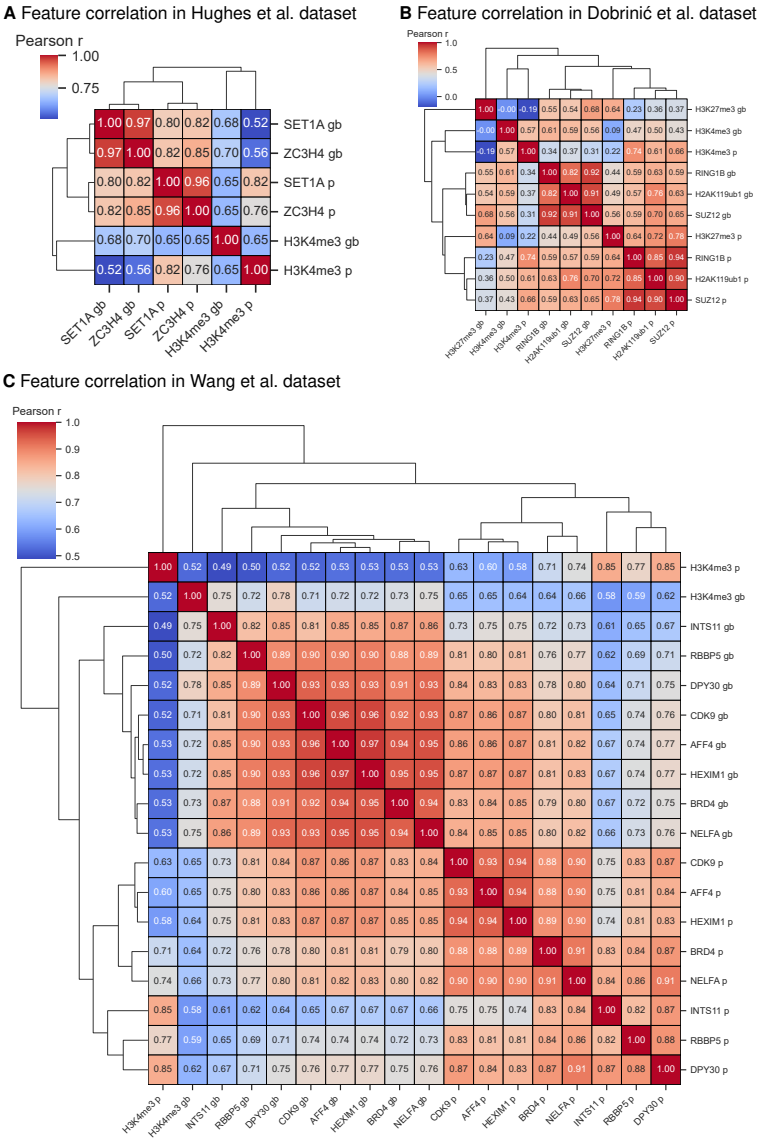
B Metrics for PyTorch MLP model trained on Dobrinić et al. data set



C Metrics for PyTorch MLP model trained on Wang et al. data set



Supplementary Methods Fig 2. Training and validation MSE losses and R^2 scores for the PyTorch MLP models trained on A) Hughes et al. [8] B) Dobrinić et al. [9] C) Wang et al. [10] data.



Supplementary Methods Fig 3. Correlation structure of chromatin features across A) Hughes et al. [8] B) Dobrinić et al. [9] and C) Wang et al. [10] data. The input data frames of training sets were used to calculate the pearson correlation coefficient matrices for all three data sets and is visualised in clustered heatmaps. Promoter=p, Gene body=gb.

References

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2;9(4):357–359. doi:10.1038/nmeth.1923.
2. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS);9(9):R137. doi:10.1186/gb-2008-9-9-r137.
3. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets;26(17):2204–2207. doi:10.1093/bioinformatics/btq351.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
5. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features;30(7):923–930. doi:10.1093/bioinformatics/btt656.
6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2;15(12):550. doi:10.1186/s13059-014-0550-8.
7. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework;. Available from: http://arxiv.org/abs/1907.10902.
8. Hughes AL, Szczurek AT, Kelley JR, Lastuvkova A, Turberfield AH, Dimitrova E, et al. A CpG island-encoded mechanism protects genes from premature transcription termination;14(1):726. doi:10.1038/s41467-023-36236-2.
9. Dobrinić P, Szczurek AT, Klose RJ. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency;28(10):811–824. doi:10.1038/s41594-021-00661-y.
10. Wang H, Fan Z, Shliaha PV, Miele M, Hendrickson RC, Jiang X, et al. H3K4me3 regulates RNA polymerase II promoter-proximal pause-release;615(7951):339–348. doi:10.1038/s41586-023-05780-8.

Source code for bioinformatics workflows and models

To ensure reproducibility and transparency, all bioinformatics workflows and machine learning models used in this thesis are publicly available. The repository includes data preprocessing pipelines, deep learning models, and SHAP-based analyses for investigating transcriptional regulation.

The repository provides:

- Bioinformatics workflows for processing RNA-seq, ChIP-seq, and TT-seq data, including alignment, peak calling, and feature quantification.
- Machine learning models for predicting RNA Pol-II occupancy and inferring regulatory interactions using SHAP-based interpretability.

The full source code is available at https://github.com/kashyapchhatbar/phdthesis

References

Abbasova, Leyla, Paulina Urbanaviciute, Di Hu, et al. (Mar. 27, 2025). "CUT&Tag recovers up to half of ENCODE ChIP-seq histone acetylation peaks". In: *Nat Commun* 16.1. Publisher: Nature Publishing Group, p. 2993. DOI: 10.1038_s41467 * 025 * 58137 * 2. (cit. on p. 23)

Alberts, Bruce (Feb. 1998). "The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists". In: *Cell* 92.3, pp. 291–294. DOI: 10.1016_S0092 * 8674.00/80922 * 8. (cit. on p. 38)

Allshire, Robin C. and Hiten D. Madhani (Apr. 2018). "Ten principles of heterochromatin formation and function". In: *Nature Reviews Molecular Cell Biology* 19.4, pp. 229–244. DOI: 10.1038_nrm.2017.119. (cit. on p. 9, 11)

Almeida, Mafalda, Greta Pintacuda, Osamu Masui, et al. (June 9, 2017). "PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation". In: *Science (New York, N.Y.)* 356.6342, pp. 1081–1084. DOI: 10.1126_science.aal2512. (cit. on p. 39)

Altschul, Stephen F., Mark S. Boguski, Warren Gish, et al. (Feb. 1, 1994). "Issues in searching molecular sequence databases". In: *Nature Genetics* 6.2, pp. 119–129. DOI: 10.1038_ng0294 * 119. (cit. on p. 24)

Amir, Ruthie E., Ignatia B. Van Den Veyver, Mimi Wan, et al. (Oct. 1999). "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2". In: *Nature Genetics* 23.2, pp. 185–188. DOI: 10.1038_13810. (cit. on p. 18, 35)

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, et al. (July 2016). "Deep learning for computational biology". In: *Molecular Systems Biology* 12.7, p. 878. DOI: 10.15252_msb.20156651. (cit. on p. 2, 25)

Baillat, David, Mohamed-Ali Hakimi, Anders M. Näär, et al. (Oct. 2005). "Integrator, a Multiprotein Mediator of Small Nuclear RNA Processing, Associates with the C-Terminal Repeat of RNA Polymerase II". In: *Cell* 123.2, pp. 265–276. DOI: 10.1016_j.cell.2005.08.019. (cit. on p. 13, 40, 41)

Banaszynski, Laura A., Ling-chun Chen, Lystranne A. Maynard-Smith, et al. (Sept. 2006). "A Rapid, Reversible, and Tunable Method to Regulate Protein Function in Living Cells Using Synthetic Small Molecules". In: *Cell* 126.5, pp. 995–1004. DOI: 10.1016_j.cell.2006.07.025. (cit. on p. 19)

Bannister, Andrew J., Robert Schneider, Fiona A. Myers, et al. (May 2005). "Spatial Distribution of Di- and Tri-methyl Lysine 36 of Histone H3 at Active Genes". In: *Journal of Biological Chemistry* 280.18, pp. 17732–17736. DOI: 10.1074_jbc.M500796200. (cit. on p. 11)

Barrios-Rodiles, Miriam, Kevin R. Brown, Barish Ozdamar, et al. (Mar. 11, 2005). "High-Throughput Mapping of a Dynamic Signaling Network in Mammalian Cells". In: *Science* 307.5715, pp. 1621–1625. DOI: 10.1126_science.1105776. (cit. on p. 38)

Barski, Artem, Suresh Cuddapah, Kairong Cui, et al. (May 2007). "High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4, pp. 823–837. DOI: 10.1016_j.cell.2007.05.009. (cit. on p. 14)

Beadle, G. W. and E. L. Tatum (Nov. 15, 1941). "Genetic Control of Biochemical Reactions in Neurospora". In: *Proceedings of the National Academy of Sciences* 27.11, pp. 499–506. DOI: 10.1073_pnas.27.11.499. (cit. on p. 38)

Bellman, Richard (2003). *Dynamic programming*. Dover ed. Mineola, N.Y: Dover Publications. 340 pp. (cit. on p. 24)

Bibikova, Marina, Mary Golic, Kent G Golic, et al. (July 1, 2002). "Targeted Chromosomal Cleavage and Mutagenesis in Drosophila Using Zinc-Finger Nucleases". In: *Genetics* 161.3, pp. 1169–1175. DOI: 10.1093_genetics_161.3.1169. (cit. on p. 19)

Bird, Adrian, Mary Taggart, Marianne Frommer, et al. (Jan. 1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA". In: *Cell* 40.1, pp. 91–99. DOI: 10.1016_0092 * 8674.85/90312 * 5. (cit. on p. 17)

Bird, Adrian P. (May 1986). "CpG-rich islands and the function of DNA methylation". In: *Nature* 321.6067, pp. 209–213. DOI: 10.1038_321209a0. (cit. on p. 37)

Bird, Adrian P. and Mary H. Taggart (1980). "Variable patterns of total DNA and rDNA methylation in animals". In: *Nucleic Acids Research* 8.7, pp. 1485–1497. DOI: 10.1093_nar_8.7.1485. (cit. on p. 16)

Blackledge, Neil P. and Robert J. Klose (Dec. 2021). "The molecular principles of gene regulation by Polycomb repressive complexes". In: *Nature Reviews Molecular Cell Biology* 22.12, pp. 815–833. DOI: 10.1038_s41580 * 021 * 00398 * y. (cit. on p. 39)

Blackledge, Neil P., Jin C. Zhou, Michael Y. Tolstorukov, et al. (Apr. 2010). "CpG Islands Recruit a Histone H3 Lysine 36 Demethylase". In: *Molecular Cell* 38.2, pp. 179–190. DOI: 10.1016_j.molcel.2010.04.009. (cit. on p. 37)

Bowen, Nathan J, Naoyuki Fujita, Masahiro Kajita, et al. (Mar. 2004). "Mi-2/NuRD: multiple complexes for many purposes". In: *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1677.1, pp. 52–57. DOI: 10.1016_j.bbaexp.2003.10.010. (cit. on p. 15)

Boxer, Lisa D., William Renthal, Alexander W. Greben, et al. (Jan. 2020). "MeCP2 Represses the Rate of Transcriptional Initiation of Highly Methylated Long Genes". In: *Molecular Cell* 77.2, 294–309.e9. DOI: 10.1016_j.molcel.2019.10.032. (cit. on p. 37)

Boyle, Alan P., Sean Davis, Hennady P. Shulha, et al. (Jan. 2008). "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2, pp. 311–322. DOI: 10.1016_j.cell.2007.12.014. (cit. on p. 23)

Bracken, Adrian P., Nikolaj Dietrich, Diego Pasini, et al. (May 1, 2006). "Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions". In: *Genes & Development* 20.9, pp. 1123–1136. DOI: 10.1101_gad.381706. (cit. on p. 14)

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023_A : 1010933404324. (cit. on p. 28)

Brenner, S., F. Jacob, and M. Meselson (May 1961). "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis". In: *Nature* 190.4776. Publisher: Nature Publishing Group, pp. 576–581. DOI: 10.1038_190576a0. (cit. on p. 5)

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, et al. (Dec. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12, pp. 1213–1218. DOI: 10.1038_nmeth.2688. (cit. on p. 23)

Chahrour, Maria, Sung Yun Jung, Chad Shaw, et al. (May 30, 2008). "MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription". In: *Science* 320.5880, pp. 1224–1229. DOI: 10.1126_science.1153252. (cit. on p. 35)

Chargaff, Erwin (1950). "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation". In: *Experientia* 6.6. Publisher: Birkhäuser-Verlag Basel, pp. 201–209. (cit. on p. 5)

Chen, Lin, Kaifu Chen, Laura A. Lavery, et al. (Apr. 28, 2015). "MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome". In: *Proceedings of the National Academy of Sciences* 112.17, pp. 5509–5514. DOI: 10.1073_pnas.1505909112. (cit. on p. 36, 58)

Chen, Xi and Hemant Ishwaran (June 2012). "Random forests for genomic data analysis". In: *Genomics* 99.6, pp. 323–329. DOI: 10.1016_j.ygeno.2012.04.003. (cit. on p. 28)

Chhatbar, Kashyap, Adrian Bird, and Guido Sanguinetti (July 15, 2025). *Modeling transcription with explainable AI uncovers context-specific epigenetic gene regulation at promoters and gene bodies*. DOI: 10.1101_2025.01.30.635704. (cit. on p. 3, 95)

Chhatbar, Kashyap, Justyna Cholewa-Waclaw, Ruth Shah, et al. (Oct. 13, 2020). "Quantitative analysis questions the role of MeCP2 as a global regulator of alternative splicing". In: *PLOS Genetics* 16.10. Ed. by Dirk Schübeler, e1009087. DOI: 10.1371_journal.pgen.1009087. (cit. on p. 3, 43)

Chhatbar, Kashyap, John Connelly, Shaun Webb, et al. (Dec. 2022). "A critique of the hypothesis that CA repeats are primary targets of neuronal MeCP2". In: *Life Science Alliance* 5.12, e202201522. DOI: 10.26508_lsa.202201522. (cit. on p. 3, 58)

Childhood Overgrowth Consortium, Katrina Tatton-Brown, Sheila Seal, et al. (Apr. 2014). "Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability". In: *Nature Genetics* 46.4, pp. 385–388. DOI: 10.1038_ng.2917. (cit. on p. 18)

Clapier, Cedric R. and Bradley R. Cairns (June 1, 2009). "The Biology of Chromatin Remodeling Complexes". In: *Annual Review of Biochemistry* 78.1, pp. 273–304. DOI: 10.1146_annurev.biochem.77.062706.153223. (cit. on p. 14, 15)

Clouaire, Thomas, Shaun Webb, Pete Skene, et al. (Jan. 8, 2012). "Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells". In: *Genes Dev.* 26.15. 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, pp. 1714–1728. DOI: 10.1101_gad.194209.112. (cit. on p. 12)

Crick, F (1958). *On protein synthesis*. Pages: 138 Publication Title: Symp Soc Exp Biol Volume: 12. (cit. on p. 5)

De Mendoza, Alex, Daniel Poppe, Sam Buckberry, et al. (Jan. 18, 2021). "The emergence of the brain non-CpG methylation system in vertebrates". In: *Nature Ecology & Evolution* 5.3, pp. 369–378. DOI: 10.1038_s41559 * 020 * 01371 * 2. (cit. on p. 18)

Deaton, Aimée M. and Adrian Bird (May 15, 2011). "CpG islands and the regulation of transcription". In: *Genes & Development* 25.10, pp. 1010–1022. DOI: 10.1101_gad.2037511. (cit. on p. 37)

Del Gaudio, Daniela, Ping Fang, Fernando Scaglia, et al. (Dec. 2006). "Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males". In: *Genetics in Medicine* 8.12, pp. 784–792. DOI: 10.1097_01.gim.0000250502.28516.3c. (cit. on p. 18)

Denissov, Sergei, Helmut Hofemeister, Hendrik Marks, et al. (Feb. 1, 2014). "Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant". In: *Development* 141.3, pp. 526–537. DOI: 10.1242_dev.102681. (cit. on p. 37)

Denslow, S A and P A Wade (Aug. 13, 2007). "The human Mi-2/NuRD complex and gene regulation". In: *Oncogene* 26.37, pp. 5433–5438. DOI: 10.1038_sj.onc.1210611. (cit. on p. 15)

Dion, Michael F., Steven J. Altschuler, Lani F. Wu, et al. (Apr. 12, 2005). "Genomic characterization reveals a simple histone H4 acetylation code". In: *Proceedings of the National Academy of Sciences* 102.15, pp. 5501–5506. DOI: 10.1073_pnas.0500136102. (cit. on p. 12)

Dobrinić, Paula, Aleksander T. Szczurek, and Robert J. Klose (Oct. 2021). "PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency". In: *Nature Structural & Molecular Biology* 28.10. Publisher: Nature Publishing Group, pp. 811–824. DOI: 10.1038_s41594 * 021 * 00661 * y. (cit. on p. 19, 40)

Eaton, Joshua D. and Steven West (Sept. 2020). "Termination of Transcription by RNA Polymerase II: BOOM!" In: *Trends in Genetics* 36.9, pp. 664–675. DOI: 10.1016_j.tig.2020.05.008. (cit. on p. 41)

Elrod, Nathan D., Telmo Henriques, Kai-Lieh Huang, et al. (Dec. 2019). "The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes". In: *Molecular Cell* 76.5, 738–752.e7. DOI: 10.1016_j.molcel.2019.10.034. (cit. on p. 40)

Erijman, Ariel, Lukasz Kozlowski, Salma Sohrabi-Jahromi, et al. (June 2020). "A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning". In: *Molecular Cell* 78.5, 890–902.e6. DOI: 10.1016_j.molcel.2020.04.020. (cit. on p. 8)

Estell, Chris, Lee Davidson, Joshua D. Eaton, et al. (July 2023). "A restrictor complex of ZC3H4, WDR82, and ARS2 integrates with PNUTS to control unproductive transcription". In: *Molecular Cell* 83.13, 2222–2239.e5. DOI: 10.1016_j.molcel.2023.05.029. (cit. on p. 13, 41)

Estell, Chris, Lee Davidson, Pieter C Steketee, et al. (Apr. 29, 2021). "ZC3H4 restricts non-coding transcription in human cells". In: *eLife* 10. Ed. by Torben Heick Jensen, James L Manley, and Joan Steitz. Publisher: eLife Sciences Publications, Ltd, e67305. DOI: 10.7554_eLife.67305. (cit. on p. 13)

Estell, Chris and Steven West (Jan. 2025). "ZC3H4/Restrictor Exerts a Stranglehold on Pervasive Transcription". In: *Journal of Molecular Biology* 437.1, p. 168707. DOI: 10.1016_j.jmb.2024.168707. (cit. on p. 13, 41)

Farcas, Anca M, Neil P Blackledge, Ian Sudbery, et al. (Dec. 18, 2012). "KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands". In: *eLife* 1, e00205. DOI: 10.7554_eLife.00205. (cit. on p. 37)

Franklin, Rosalind E. and R. G. Gosling (Apr. 25, 1953). "Molecular Configuration in Sodium Thymonucleate". In: *Nature* 171.4356, pp. 740–741. DOI: 10.1038_171740a0. (cit. on p. 5)

Frietze, Seth and Peggy J. Farnham (2011). "Transcription Factor Effector Domains". In: *A Handbook of Transcription Factors*. Ed. by Timothy R. Hughes. Vol. 52. Series

Title: Subcellular Biochemistry. Dordrecht: Springer Netherlands, pp. 261–277. DOI: 10.1007_978 * 90 * 481 * 9069 * 0_12. (cit. on p. 8)

Frommer, M, L E McDonald, D S Millar, et al. (Mar. 1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." In: *Proceedings of the National Academy of Sciences* 89.5, pp. 1827–1831. DOI: 10.1073_pnas.89.5.1827. (cit. on p. 23)

Furey, Terrence S. (Dec. 2012). "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions". In: *Nature Reviews Genetics* 13.12, pp. 840–852. DOI: 10.1038_nrg3306. (cit. on p. 22)

Gabel, Harrison W., Benyam Kinde, Hume Stroud, et al. (June 2015). "Disruption of DNA-methylation-dependent long gene repression in Rett syndrome". In: *Nature* 522.7554, pp. 89–93. DOI: 10.1038_nature14319. (cit. on p. 36, 58)

Gao, Chong, Nikki R. Kong, and Li Chai (2011). "The Role of Stem Cell Factor SALL4 in Leukemogenesis". In: *Critical Reviews™ in Oncogenesis* 16.1, pp. 117–127. DOI: 10.1615_CritRevOncog.v16.i1 * 2.110. (cit. on p. 38)

Gilbert, Walter (Feb. 1978). "Why genes in pieces?" In: *Nature* 271.5645, pp. 501–501. DOI: 10.1038_271501a0. (cit. on p. 6)

Gingras, Anne-Claude, Matthias Gstaiger, Brian Raught, et al. (Aug. 2007). "Analysis of protein complexes using mass spectrometry". In: *Nature Reviews Molecular Cell Biology* 8.8, pp. 645–654. DOI: 10.1038_nrm2208. (cit. on p. 38)

Greenberg, Maxim V. C. and Deborah Bourc'his (Oct. 2019). "The diverse roles of DNA methylation in mammalian development and disease". In: *Nature Reviews Molecular Cell Biology* 20.10, pp. 590–607. DOI: 10.1038_s41580 * 019 * 0159 * 6. (cit. on p. 16)

Gros, Francois, H. Hiatt, Walter Gilbert, et al. (May 1961). "Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli". In: *Nature* 190.4776. Publisher: Nature Publishing Group, pp. 581–585. DOI: 10.1038_190581a0. (cit. on p. 5)

Guenther, M. G., W. S. Lane, W. Fischle, et al. (May 1, 2000). "A core SMRT corepressor complex containing HDAC3 and TBL1, a WD40-repeat protein linked to deafness". In: *Genes & Development* 14.9, pp. 1048–1057. (cit. on p. 15)

Guo, Junjie U, Yijing Su, Joo Heon Shin, et al. (Feb. 2014). "Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain". In: *Nature Neuroscience* 17.2, pp. 215–222. DOI: 10.1038_nn.3607. (cit. on p. 58)

Guy, Jacky, Jian Gan, Jim Selfridge, et al. (Feb. 23, 2007). "Reversal of neurological defects in a mouse model of Rett syndrome". In: *Science (New York, N.Y.)* 315.5815, pp. 1143–1147. DOI: 10.1126_science.1138389. (cit. on p. 125)

Hamagami, Nicole, Dennis Y. Wu, Adam W. Clemens, et al. (May 2023). "NSD1 deposits histone H3 lysine 36 dimethylation to pattern non-CG DNA methylation in neurons". In: *Molecular Cell* 83.9, 1412–1428.e7. DOI: 10.1016_j.molcel.2023.04.001. (cit. on p. 18)

Hashimoto, Hideharu, Yiwei Liu, Anup K. Upadhyay, et al. (June 2012). "Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation". In: *Nucleic Acids Research* 40.11, pp. 4841–4849. DOI: 10.1093_nar_gks155. (cit. on p. 17)

He, Yupeng and Joseph R. Ecker (Aug. 24, 2015). "Non-CG Methylation in the Human Genome". In: *Annual Review of Genomics and Human Genetics* 16.1, pp. 55–77. DOI: 10.1146_annurev * genom * 090413 * 025437. (cit. on p. 17, 36)

Hendrich, Brian and Adrian Bird (Nov. 1, 1998). "Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins". In: *Molecular and Cellular Biology* 18.11, pp. 6538–6547. DOI: 10.1128_MCB.18.11.6538. (cit. on p. 17)

Henikoff, Steven and Ali Shilatifard (Oct. 2011). "Histone modification: cause or cog?" In: *Trends in Genetics* 27.10, pp. 389–396. DOI: 10.1016_j.tig.2011.06.006.(cit. on p. 10)

Herzog, Veronika A, Brian Reichholf, Tobias Neumann, et al. (Dec. 1, 2017). "Thiol-linked alkylation of RNA to assess expression dynamics". In: *Nature Methods* 14.12, pp. 1198–1204. DOI: 10.1038_nmeth.4435. (cit. on p. 22)

Heskes, Tom, Evi Sijben, Ioan Gabriel Bucur, et al. (Nov. 3, 2020). *Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models*. DOI: 10.48550_arXiv.2011.01625. arXiv: 2011.01625[cs]. (cit. on p. 126)

Ho, Kok Lian, Iain W. McNae, Lars Schmiedeberg, et al. (Feb. 2008). "MeCP2 Binding to DNA Depends upon Hydration at Methyl-CpG". In: *Molecular Cell* 29.4, pp. 525–531. DOI: 10.1016_j.molcel.2007.12.028. (cit. on p. 35–37)

Holland, Andrew J., Daniele Fachinetti, Joo Seok Han, et al. (Dec. 4, 2012). "Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells". In: *Proceedings of the National Academy of Sciences* 109.49. DOI: 10.1073_pnas.1216880109. (cit. on p. 19)

Holliday, R. and G.W. Grigg (Jan. 1993). "DNA methylation and mutation". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 285.1, pp. 61–67. DOI: 10.1016_0027 * 5107.93/90052 * H. (cit. on p. 16, 18)

Holliday, R. and J. E. Pugh (Jan. 24, 1975). "DNA modification mechanisms and gene activity during development". In: *Science (New York, N.Y.)* 187.4173, pp. 226–232.(cit. on p. 17)

Hotchkiss, Rollin D. (Aug. 1948). "THE QUANTITATIVE SEPARATION OF PURINES, PYRIMIDINES, AND NUCLEOSIDES BY PAPER CHROMATOGRAPHY". In: *Journal of Biological Chemistry* 175.1, pp. 315–332. DOI: 10.1016_S0021 * 9258.18/57261 * 6.(cit. on p. 16)

Hughes, Amy L., Jessica R. Kelley, and Robert J. Klose (Aug. 1, 2020). "Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.8, p. 194567. DOI: 10.1016_j.bbagrm.2020.194567. (cit. on p. 11, 12)

Hughes, Amy L., Aleksander T. Szczurek, Jessica R. Kelley, et al. (Feb. 9, 2023). "A CpG island-encoded mechanism protects genes from premature transcription termination". In: *Nature Communications* 14.1. Publisher: Nature Publishing Group, p. 726. DOI: 10.1038_s41467 * 023 * 36236 * 2. (cit. on p. 19, 41, 42)

Husmann, Dylan and Or Gozani (Oct. 2019). "Histone lysine methyltransferases in biology and disease". In: *Nature Structural & Molecular Biology* 26.10, pp. 880–889. DOI: 10.1038_s41594 * 019 * 0298 * 7. (cit. on p. 11)

Ibrahim, Abdulkhaleg, Christophe Papin, Kareem Mohideen-Abdul, et al. (June 25, 2021). "MeCP2 is a microsatellite binding protein that protects CA repeats from nucleosome invasion". In: *Science* 372.6549, eabd5581. DOI: 10.1126_science.abd5581. (cit. on p. 35, 37, 58)

Illingworth, Robert S., Ulrike Gruenewald-Schneider, Shaun Webb, et al. (Sept. 23, 2010). "Orphan CpG Islands Identify Numerous Conserved Promoters in the

Mammalian Genome". In: *PLoS Genetics* 6.9. Ed. by Wolf Reik, e1001134. DOI: 10.1371_journal.pgen.1001134. (cit. on p. 18)

Ishiyama, Satoshi, Atsuya Nishiyama, Yasushi Saeki, et al. (Oct. 2017). "Structure of the Dnmt1 Reader Module Complexed with a Unique Two-Mono-Ubiquitin Mark on Histone H3 Reveals the Basis for DNA Methylation Maintenance". In: *Molecular Cell* 68.2, 350–360.e7. DOI: 10.1016_j.molcel.2017.09.037. (cit. on p. 17)

Ito, Shinsuke, Ana C. D'Alessio, Olena V. Taranova, et al. (Aug. 2010). "Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification". In: *Nature* 466.7310, pp. 1129–1133. DOI: 10.1038_nature09303.(cit. on p. 17)

Ito, Shinsuke, Li Shen, Qing Dai, et al. (Sept. 2, 2011). "Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine". In: *Science* 333.6047, pp. 1300–1303. DOI: 10.1126_science.1210597. (cit. on p. 17)

Jacob, François and Jacques Monod (June 1, 1961). "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3, pp. 318–356. DOI: 10.1016_S0022 * 2836.61/80072 * 7. (cit. on p. 5)

Jaeger, Martin G. and Georg E. Winter (Apr. 2021). "Fast-acting chemical tools to delineate causality in transcriptional control". In: *Molecular Cell* 81.8, pp. 1617–1630. DOI: 10.1016_j.molcel.2021.02.015. (cit. on p. 2, 19)

Jaenisch, Rudolf and Adrian Bird (Mar. 2003). "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals". In: *Nature Genetics* 33 (S3), pp. 245–254. DOI: 10.1038_ng1089. (cit. on p. 16)

Jain, Ankur, Ruijie Liu, Biswarathan Ramani, et al. (May 2011). "Probing cellular protein complexes using single-molecule pull-down". In: *Nature* 473.7348, pp. 484–488. DOI: 10.1038_nature10016. (cit. on p. 38)

Jiang, Yan, Xing Fu, Yuhan Zhang, et al. (Oct. 1, 2021). "Rett syndrome linked to defects in forming the MeCP2/Rbfox/LASR complex in mouse models". In: *Nature Communications* 12.1, p. 5767. DOI: 10.1038_s41467 * 021 * 26084 * 3. (cit. on p. 125)

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, et al. (Aug. 17, 2012). "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity". In: *Science* 337.6096, pp. 816–821. DOI: 10.1126_science.1225829. (cit. on p. 19)

Johnson, Treat B. and Robert D. Coghill (Nov. 1925). "RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS[1]". In: *Journal of the American Chemical Society* 47.11, pp. 2838–2844. DOI: 10.1021_ja01688a030. (cit. on p. 16)

Jolma, Arttu, Jian Yan, Thomas Whitington, et al. (Jan. 2013). "DNA-Binding Specificities of Human Transcription Factors". In: *Cell* 152.1, pp. 327–339. DOI: 10.1016_j.cell.2012.12.009. (cit. on p. 8)

Jones, Peter A. (July 2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7, pp. 484–492. DOI: 10.1038_nrg3230. (cit. on p. 16)

Kankirawatana, P., H. Leonard, C. Ellaway, et al. (July 11, 2006). "Early progressive encephalopathy in boys and *MECP2* mutations". In: *Neurology* 67.1, pp. 164–166. DOI: 10.1212_01.wnl.0000223318.28938.45. (cit. on p. 35)

Kaya-Okur, Hatice S., Steven J. Wu, Christine A. Codomo, et al. (Apr. 29, 2019). "CUT&Tag for efficient epigenomic profiling of small samples and single cells". In: *Nat Commun* 10.1. Publisher: Nature Publishing Group, p. 1930. DOI: 10.1038_s41467 * 019 * 09982 * 5. (cit. on p. 22)

Kharchenko, Peter V., Artyom A. Alekseyenko, Yuri B. Schwartz, et al. (Mar. 2011). "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster". In: *Nature* 471.7339, pp. 480–485. DOI: 10.1038_nature09725. (cit. on p. 12)

Kiefer, Susan McLeskey, Bradley W. McDill, Jing Yang, et al. (Apr. 2002). "Murine Sall1 Represses Transcription by Recruiting a Histone Deacetylase Complex". In: *Journal of Biological Chemistry* 277.17, pp. 14869–14876. DOI: 10.1074_jbc.M200052200. (cit. on p. 15)

Kim, Hyongbum and Jin-Soo Kim (May 2014). "A guide to genome engineering with programmable nucleases". In: *Nature Reviews Genetics* 15.5, pp. 321–334. DOI: 10.1038_nrg3686. (cit. on p. 18)

Kim, Jeongkyu, Mingeun Ji, and Gangman Yi (2020). "A Review on Sequence Alignment Algorithms for Short Reads Based on Next-Generation Sequencing". In: *IEEE Access* 8, pp. 189811–189822. DOI: 10.1109_ACCESS.2020.3031159. (cit. on p. 21)

Kinde, Benyam, Dennis Y. Wu, Michael E. Greenberg, et al. (Dec. 27, 2016). "DNA methylation in the gene body influences MeCP2-mediated gene repression". In: *Proceedings of the National Academy of Sciences* 113.52, pp. 15114–15119. DOI: 10.1073_pnas.1618737114. (cit. on p. 36)

Kizer, Kelby O., Hemali P. Phatnani, Yoichiro Shibata, et al. (Apr. 1, 2005). "A Novel Domain in Set2 Mediates RNA Polymerase II Interaction and Couples Histone H3 K36 Methylation with Transcript Elongation". In: *Molecular and Cellular Biology* 25.8, pp. 3305–3316. DOI: 10.1128_MCB.25.8.3305 * 3316.2005. (cit. on p. 11)

Klose, Robert J. and Adrian P. Bird (Feb. 2006). "Genomic DNA methylation: the mark and its mediators". In: *Trends in Biochemical Sciences* 31.2, pp. 89–97. DOI: 10.1016_j.tibs.2005.12.008. (cit. on p. 16)

Klose, Robert J., Sarah Cooper, Anca M. Farcas, et al. (Aug. 22, 2013). "Chromatin Sampling—An Emerging Perspective on Targeting Polycomb Repressor Proteins". In: *PLOS Genetics* 9.8. Publisher: Public Library of Science, e1003717. DOI: 10.1371_journal.pgen.1003717. (cit. on p. 13)

Kohlhase, Jürgen, Annegret Wischermann, Herbert Reichenbach, et al. (Jan. 1998). "Mutations in the SALL1 putative transcription factor gene cause Townes-Brocks syndrome". In: *Nature Genetics* 18.1, pp. 81–83. DOI: 10.1038_ng0198 * 81. (cit. on p. 38)

Kornberg, R D (Dec. 2007). "The molecular basis of eucaryotic transcription". In: *Cell Death & Differentiation* 14.12, pp. 1989–1997. DOI: 10.1038_sj.cdd.4402251. (cit. on p. 6)

Köster, Johannes and Sven Rahmann (Oct. 1, 2012). "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19, pp. 2520–2522. DOI: 10.1093_bioinformatics_bts480. (cit. on p. 25)

Kriaucionis, Skirmantas and Nathaniel Heintz (May 15, 2009). "The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain". In: *Science* 324.5929, pp. 929–930. DOI: 10.1126_science.1169786. (cit. on p. 17)

Kurdistani, Siavash K, Saeed Tavazoie, and Michael Grunstein (June 2004). "Mapping Global Histone Acetylation Patterns to Gene Expression". In: *Cell* 117.6, pp. 721–733. DOI: 10.1016_j.cell.2004.05.023. (cit. on p. 12)

Kurotaki, Naohiro, Kiyoshi Imaizumi, Naoki Harada, et al. (Apr. 2002). "Haploinsufficiency of NSD1 causes Sotos syndrome". In: *Nature Genetics* 30.4, pp. 365–366. DOI: 10.1038_ng863. (cit. on p. 18)

Lagger, Sabine, John C. Connelly, Gabriele Schweikert, et al. (May 12, 2017). "MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain". In: *PLOS Genetics* 13.5. Ed. by John M. Greally, e1006793. DOI: 10.1371_journal.pgen.1006793. (cit. on p. 36, 37, 58)

Lai, Anne Y. and Paul A. Wade (Aug. 2011). "Cancer biology and NuRD: a multifaceted chromatin remodelling complex". In: *Nature Reviews Cancer* 11.8, pp. 588–596. DOI: 10.1038_nrc3091. (cit. on p. 15)

Lai, Fan, Alessandro Gardini, Anda Zhang, et al. (Sept. 2015). "Integrator mediates the biogenesis of enhancer RNAs". In: *Nature* 525.7569, pp. 399–403. DOI: 10.1038_nature14906. (cit. on p. 41)

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, et al. (Feb. 2018). "The Human Transcription Factors". In: *Cell* 172.4, pp. 650–665. DOI: 10.1016_j.cell.2018.01.029. (cit. on p. 8)

Lauberth, Shannon M. and Michael Rauchman (Aug. 2006). "A Conserved 12-Amino Acid Motif in Sall1 Recruits the Nucleosome Remodeling and Deacetylase Corepressor Complex". In: *Journal of Biological Chemistry* 281.33, pp. 23922–23931. DOI: 10.1074_jbc.M513461200. (cit. on p. 15, 38)

Lavery, Laura A, Kerstin Ure, Ying-Wooi Wan, et al. (Mar. 11, 2020). "Losing Dnmt3a dependent methylation in inhibitory neurons impairs neural function by a mechanism impacting Rett syndrome". In: *eLife* 9, e52981. DOI: 10.7554_eLife.52981. (cit. on p. 36, 58)

Layden, Hillary M., Anna E. Johnson, and Scott W. Hiebert (Jan. 2024). "Chemical-genetics refines transcription factor regulatory circuits". In: *Trends in Cancer* 10.1, pp. 65–75. DOI: 10.1016_j.trecan.2023.08.012. (cit. on p. 20)

Lee, Jeong-Heon, Jinsam You, Erika Dobrota, et al. (Aug. 6, 2010). "Identification and Characterization of a Novel Human PP1 Phosphatase Complex *". In: *Journal of Biological Chemistry* 285.32. Publisher: Elsevier, pp. 24466–24476. DOI: 10.1074_jbc.M110.109801. (cit. on p. 13)

Legube, Gaëlle and Didier Trouche (Oct. 2003). "Regulating histone acetyltransferases and deacetylases". In: *EMBO reports* 4.10, pp. 944–947. DOI: 10.1038_sj.embor.embor941. (cit. on p. 14)

Lev Maor, Galit, Ahuvi Yearim, and Gil Ast (May 2015). "The alternative role of DNA methylation in splicing regulation". In: *Trends in Genetics* 31.5, pp. 274–280. DOI: 10.1016_j.tig.2015.03.002. (cit. on p. 37)

Levene, P.A. (Dec. 1919). "THE STRUCTURE OF YEAST NUCLEIC ACID". In: *Journal of Biological Chemistry* 40.2, pp. 415–424. DOI: 10.1016_S0021 * 9258.18/87254 * 4. (cit. on p. 5)

Lewis, Joe D., Richard R. Meehan, William J. Henzel, et al. (June 1992). "Purification, sequence, and cellular localization of a novel chromosomal protein that binds to

Methylated DNA". In: *Cell* 69.6, pp. 905–914. DOI: 10.1016_0092 * 8674.92/90610 * O. (cit. on p. 35)

Li, J. (Aug. 15, 2000). "Both corepressor proteins SMRT and N-CoR exist in large protein complexes containing HDAC3". In: *The EMBO Journal* 19.16, pp. 4342–4350. DOI: 10.1093_emboj_19.16.4342. (cit. on p. 15)

Libbrecht, Maxwell W. and William Stafford Noble (June 2015). "Machine learning applications in genetics and genomics". In: *Nature Reviews Genetics* 16.6, pp. 321–332. DOI: 10.1038_nrg3920. (cit. on p. 28)

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (Dec. 25, 2020). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1, p. 18. DOI: 10.3390_e23010018. (cit. on p. 31)

Ling, X, T A Harkness, M C Schultz, et al. (Mar. 15, 1996). "Yeast histone H3 and H4 amino termini are important for nucleosome assembly in vivo and in vitro: redundant and position-independent functions in assembly but not in gene regulation." In: *Genes & Development* 10.6, pp. 686–699. DOI: 10.1101_gad.10.6.686. (cit. on p. 12)

Lister, Ryan and Joseph R. Ecker (June 2009). "Finding the fifth base: Genome-wide sequencing of cytosine methylation". In: *Genome Research* 19.6, pp. 959–966. DOI: 10.1101_gr.083451.108. (cit. on p. 23)

Lister, Ryan, Eran A. Mukamel, Joseph R. Nery, et al. (Aug. 9, 2013). "Global Epigenomic Reconfiguration During Mammalian Brain Development". In: *Science* 341.6146, p. 1237905. DOI: 10.1126_science.1237905. (cit. on p. 36)

Liu, Yi, Anthony Flamier, George W. Bell, et al. (June 2024). "MECP2 directly interacts with RNA polymerase II to modulate transcription in human neurons". In: *Neuron* 112.12, 1943–1958.e10. DOI: 10.1016_j.neuron.2024.04.007. (cit. on p. 125)

Lu, Jiayun, Hawon Jeong, Nikki Kong, et al. (May 18, 2009). "Stem Cell Factor SALL4 Represses the Transcriptions of PTEN and SALL1 through an Epigenetic Repressor Complex". In: *PLoS ONE* 4.5. Ed. by Maurizio Capogrossi, e5577. DOI: 10.1371_journal.pone.0005577. (cit. on p. 15)

Luger, Karolin, Armin W. Mäder, Robin K. Richmond, et al. (Sept. 1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution". In: *Nature* 389.6648, pp. 251–260. DOI: 10.1038_38444. (cit. on p. 9)

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc.(cit. on p. 25, 31, 32, 39)

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee (Mar. 7, 2019). *Consistent Individualized Feature Attribution for Tree Ensembles*. DOI: 10.48550_arXiv.1802.03888. arXiv: 1802.03888[cs]. (cit. on p. 33)

Lykke-Andersen, Søren, Kristina Žumer, Ewa Šmidová Molska, et al. (Feb. 2021). "Integrator is a genome-wide attenuator of non-productive transcription". In: *Molecular Cell* 81.3, 514–529.e6. DOI: 10.1016_j.molcel.2020.12.014. (cit. on p. 41)

Lyst, Matthew J, Robert Ekiert, Daniel H Ebert, et al. (July 2013). "Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor". In: *Nature Neuroscience* 16.7, pp. 898–902. DOI: 10.1038_nn.3434. (cit. on p. 16, 18, 35)

Lyst, Matthew J. and Adrian Bird (May 2015). "Rett syndrome: a complex disorder with simple roots". In: *Nature Reviews Genetics* 16.5, pp. 261–275. DOI: 10.1038_nrg3897. (cit. on p. 35)

Macrae, Trisha A., Julie Fothergill-Robinson, and Miguel Ramalho-Santos (Jan. 2023). "Regulation, functions and transmission of bivalent chromatin during mammalian development". In: *Nature Reviews Molecular Cell Biology* 24.1. Publisher: Nature Publishing Group, pp. 6–26. DOI: 10.1038_s41580 * 022 * 00518 * 2. (cit. on p. 40)

Marasco, Luciano E. and Alberto R. Kornblihtt (Apr. 2023). "The physiology of alternative splicing". In: *Nature Reviews Molecular Cell Biology* 24.4, pp. 242–254. DOI: 10.1038_s41580 * 022 * 00545 * z. (cit. on p. 6, 7)

Mariuswalter (July 5, 2015). *English: Schematic representation of histone modifications. Based on Rodriguez-Paredes and Esteller, Nature, 2011.* (cit. on p. 11)

Markowetz, Florian (Mar. 9, 2017). "All biology is computational biology". In: *PLOS Biology* 15.3, e2002050. DOI: 10.1371_journal.pbio.2002050. (cit. on p. 1)

Marx, Vivien (June 13, 2013). "The big challenges of big data". In: *Nature* 498.7453, pp. 255–260. DOI: 10.1038_498255a. (cit. on p. 1)

McCarty, M. and O. T. Avery (Jan. 31, 1946). "STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : II. EFFECT OF DESOXYRIBONUCLEASE ON THE BIOLOGICAL ACTIVITY OF THE TRANSFORMING SUBSTANCE". In: *The Journal of Experimental Medicine* 83.2, pp. 89–96. (cit. on p. 5)

Meehan, Richard R., Joe D. Lewis, Stewart McKay, et al. (Aug. 1989). "Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs". In: *Cell* 58.3, pp. 499–507. DOI: 10.1016_0092 * 8674.89/90430 * 3. (cit. on p. 17)

Megee, P C, B A Morgan, and M M Smith (July 15, 1995). "Histone H4 and the maintenance of genome integrity." In: *Genes & Development* 9.14, pp. 1716–1727. DOI: 10.1101_gad.9.14.1716. (cit. on p. 12)

Meneghini, Marc D., Michelle Wu, and Hiten D. Madhani (Mar. 2003). "Conserved Histone Variant H2A.Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin". In: *Cell* 112.5, pp. 725–736. DOI: 10.1016_S0092 * 8674.03/00123 * 5. (cit. on p. 9)

Meyer, Clifford A. and X. Shirley Liu (Nov. 2014). "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology". In: *Nature Reviews Genetics* 15.11, pp. 709–721. DOI: 10.1038_nrg3788. (cit. on p. 22)

Miescher, Johann Friedrich (1871). *Ueber die chemische Zusammensetzung der Eiterzellen.* (cit. on p. 5)

Mikkelsen, Tarjei S., Manching Ku, David B. Jaffe, et al. (Aug. 2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153, pp. 553–560. DOI: 10.1038_nature06008. (cit. on p. 14)

Millán-Zambrano, Gonzalo, Adam Burton, Andrew J. Bannister, et al. (Sept. 2022). "Histone post-translational modifications — cause and consequence of genome function". In: *Nature Reviews Genetics* 23.9, pp. 563–580. DOI: 10.1038_s41576 * 022 * 00468 * 7. (cit. on p. 9, 10)

Mimoso, Claudia A., Hanneke Vlaming, Nathalie P. De Wagenaar, et al. (Jan. 10, 2025). *Restrictor slows early transcription elongation to render RNA polymerase II susceptible to termination at non-coding RNA loci.* DOI: 10.1101_2025.01.08.631787.(cit. on p. 124)

Mishra, Gyan Prakash, Eric X. Sun, Tiffany Chin, et al. (Jan. 2025). "Interaction of methyl-CpG-binding protein 2 (MeCP2) with distinct enhancers in the mouse cortex". In: *Nature Neuroscience* 28.1, pp. 62–71. DOI: 10.1038_s41593 * 024 * 01808 * y. (cit. on p. 35, 125)

Molnar, Christoph (2022). *Interpretable machine learning: a guide for making black box models explainable*. Second edition. Munich, Germany: Christoph Molnar. 1 p. (cit. on p. 2, 25)

Murphy, Kevin P. (2023). *Probabilistic machine learning: advanced topics*. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press. 1 p. (cit. on p. 29)

Nakagawa, Shinichi and Holger Schielzeth (Nov. 2010). "Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists". In: *Biological Reviews* 85.4, pp. 935–956. DOI: 10.1111_j.1469 * 185X.2010.00141.x. (cit. on p. 26)

Nan, Xinsheng, Huck-Hui Ng, Colin A. Johnson, et al. (May 1998). "Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex". In: *Nature* 393.6683, pp. 386–389. DOI: 10.1038_30764.(cit. on p. 15, 17, 35)

Needleman, Saul B. and Christian D. Wunsch (Mar. 1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3, pp. 443–453. DOI: 10.1016_0022 * 2836.70/90057 * 4.(cit. on p. 24)

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, p. 370. DOI: 10.2307_2344614. (cit. on p. 26)

Nepusz, Tamás, Haiyuan Yu, and Alberto Paccanaro (May 2012). "Detecting overlapping protein complexes in protein-protein interaction networks". In: *Nature Methods* 9.5, pp. 471–472. DOI: 10.1038_nmeth.1938. (cit. on p. 38)

Neugebauer, Karla M. (Oct. 15, 2002). "On the importance of being co-transcriptional". In: *Journal of Cell Science* 115.20, pp. 3865–3871. DOI: 10.1242_jcs.00073.(cit. on p. 7)

Ng, Huck-Hui, Yi Zhang, Brian Hendrich, et al. (Sept. 1999). "MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex". In: *Nature Genetics* 23.1, pp. 58–61. DOI: 10.1038_12659. (cit. on p. 17)

Nicetto, Dario, Greg Donahue, Tanya Jain, et al. (Jan. 18, 2019). "H3K9me3-heterochromatin loss at protein-coding genes enables developmental lineage specification". In: *Science* 363.6424, pp. 294–297. DOI: 10.1126_science.aau0583. (cit. on p. 11)

Nishiyama, Atsuya, Luna Yamaguchi, Jafar Sharif, et al. (Oct. 2013). "Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication". In: *Nature* 502.7470, pp. 249–253. DOI: 10.1038_nature12488. (cit. on p. 17)

Okano, Masaki, Daphne W Bell, Daniel A Haber, et al. (Oct. 1999). "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development". In: *Cell* 99.3, pp. 247–257. DOI: 10.1016_S0092 * 8674.00/81656 * 6. (cit. on p. 17)

Okano, Masaki, Shaoping Xie, and En Li (July 1998). "Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases". In: *Nature Genetics* 19.3, pp. 219–220. DOI: 10.1038_890. (cit. on p. 17)

Orphanides, George and Danny Reinberg (Feb. 2002). "A Unified Theory of Gene Expression". In: *Cell* 108.4, pp. 439–451. DOI: 10.1016_S0092 * 8674.02/00655 * 4. (cit. on p. 6)

Otani, Junji, Hironobu Kimura, Jafar Sharif, et al. (Dec. 10, 2013). "Cell Cycle-Dependent Turnover of 5-Hydroxymethyl Cytosine in Mouse Embryonic Stem Cells". In: *PLoS ONE* 8.12. Ed. by Guoliang Xu, e82961. DOI: 10.1371_journal.pone.0082961.(cit. on p. 17)

Ou, Horng D., Sébastien Phan, Thomas J. Deerinck, et al. (July 28, 2017). "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells". In: *Science* 357.6349, eaag0025. DOI: 10.1126_science.aag0025. (cit. on p. 10)

Pabinger, S., A. Dander, M. Fischer, et al. (Mar. 1, 2014). "A survey of tools for variant analysis of next-generation genome sequencing data". In: *Briefings in Bioinformatics* 15.2, pp. 256–278. DOI: 10.1093_bib_bbs086. (cit. on p. 2)

Palmieri, Michela, Diego Pozzer, and Nicoletta Landsberger (May 25, 2023). "Advanced genetic therapies for the treatment of Rett syndrome: state of the art and future perspectives". In: *Frontiers in Neuroscience* 17, p. 1172805. DOI: 10.3389_fnins.2023.1172805. (cit. on p. 126)

Pantier, Raphaël, Megan Brown, Sicheng Han, et al. (May 8, 2024). "MeCP2 binds to methylated DNA independently of phase separation and heterochromatin organisation". In: *Nature Communications* 15.1, p. 3880. DOI: 10.1038_s41467 * 024 * 47395 * 1. (cit. on p. 68)

Pantier, Raphaël, Kashyap Chhatbar, Timo Quante, et al. (Feb. 2021). "SALL4 controls cell fate in response to DNA base composition". In: *Molecular Cell* 81.4, 845–858.e8. DOI: 10.1016_j.molcel.2020.11.046. (cit. on p. 3, 69)

Papin, Christophe, Abdulkhaleg Ibrahim, Stéphanie Le Gras, et al. (June 2017). "Combinatorial DNA methylation codes at repetitive elements". In: *Genome Research* 27.6, pp. 934–946. DOI: 10.1101_gr.213983.116. (cit. on p. 58)

Pearl, Judea (Oct. 16, 2012). *The Do-Calculus Revisited*. DOI: 10.48550_arXiv.1210.4852. arXiv: 1210.4852[cs]. (cit. on p. 126)

Perissi, Valentina, Kristen Jepsen, Christopher K. Glass, et al. (Feb. 2010). "Deconstructing repression: evolving models of co-repressor action". In: *Nature Reviews Genetics* 11.2, pp. 109–123. DOI: 10.1038_nrg2736. (cit. on p. 14, 15)

Plackett, R. L. (1972). "Studies in the History of Probability and Statistics. XXIX: The discovery of the method of least squares". In: *Biometrika* 59.2, pp. 239–251. DOI: 10.1093_biomet_59.2.239. (cit. on p. 26)

Policarpi, Cristina, Marzia Munafò, Stylianos Tsagkris, et al. (June 2024). "Systematic epigenome editing captures the context-dependent instructive function of chromatin modifications". In: *Nature Genetics* 56.6, pp. 1168–1180. DOI: 10.1038_s41588 * 024 * 01706 * w. (cit. on p. 10)

Proudfoot, Nick J., Andre Furger, and Michael J. Dye (Feb. 2002). "Integrating mRNA Processing with Transcription". In: *Cell* 108.4, pp. 501–512. DOI: 10.1016_S0092 * 8674.02/00617 * 7. (cit. on p. 6)

Puig, Oscar, Friederike Caspary, Guillaume Rigaut, et al. (July 2001). "The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification". In: *Methods* 24.3, pp. 218–229. DOI: 10.1006_meth.2001.1183. (cit. on p. 38)

Qin, Weihua, Patricia Wolf, Nan Liu, et al. (Aug. 2015). “DNA methylation requires a DNMT1 ubiquitin interacting motif (UIM) and histone ubiquitination”. In: *Cell Research* 25.8, pp. 911–929. DOI: 10.1038_cr.2015.72. (cit. on p. 17)

Quante, Timo and Adrian Bird (Apr. 2016). “Do short, frequent DNA sequence motifs mould the epigenome?” In: *Nature Reviews Molecular Cell Biology* 17.4, pp. 257–262. DOI: 10.1038_nrm.2015.31. (cit. on p. 38)

Raman, Ayush T., Amy E. Pohodich, Ying-Wooi Wan, et al. (Aug. 13, 2018). “Apparent bias toward long gene misregulation in MeCP2 syndromes disappears after controlling for baseline variations”. In: *Nature Communications* 9.1, p. 3225. DOI: 10.1038_s41467*018 * 05627 * 1. (cit. on p. 36)

Reed, Robin and Ed Hurt (Feb. 2002). “A Conserved mRNA Export Machinery Coupled to pre-mRNA Splicing”. In: *Cell* 108.4, pp. 523–531. DOI: 10.1016_S0092*8674.02/00627*X. (cit. on p. 6)

Reichard, Julia and Geraldine Zimmer-Bensch (Nov. 3, 2021). “The Epigenome in Neurodevelopmental Disorders”. In: *Frontiers in Neuroscience* 15, p. 776809. DOI: 10.3389_fnins.2021.776809. (cit. on p. 18)

Reiter, Franziska, Sebastian Wienerroither, and Alexander Stark (Apr. 2017). “Combinatorial function of transcription factors and cofactors”. In: *Current Opinion in Genetics & Development* 43, pp. 73–81. DOI: 10.1016_j.gde.2016.12.007. (cit. on p. 8)

Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder (May 2015). “High-Throughput Sequencing Technologies”. In: *Molecular Cell* 58.4, pp. 586–597. DOI: 10.1016_j.molcel.2015.05.004. (cit. on p. 21)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 9, 2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* DOI: 10.48550_arXiv.1602.04938. arXiv: 1602.04938[cs]. (cit. on p. 33)

Rideout, William M., Gerhard A. Coetzee, Aria F. Olumi, et al. (Sept. 14, 1990). “5-Methylcytosine as an Endogenous Mutagen in the Human LDL Receptor and p53 Genes”. In: *Science* 249.4974, pp. 1288–1290. DOI: 10.1126_science.1697983. (cit. on p. 16, 18)

Riggs, A.D. (1975). “X inactivation, differentiation, and DNA methylation”. In: *Cytogenetic and Genome Research* 14.1, pp. 9–25. DOI: 10.1159_000130315. (cit. on p. 17)

Robbins, Herbert and Sutton Monro (Sept. 1951). “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407. DOI: 10.1214_aoms_1177729586. (cit. on p. 30)

Rodríguez-Paredes, Manuel and Manel Esteller (Mar. 2011). “Cancer epigenetics reaches mainstream oncology”. In: *Nature Medicine* 17.3, pp. 330–339. DOI: 10.1038_nm.2305. (cit. on p. 11)

Rosenblatt, Frank (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Google-Books-ID: 7FhRAAAAMAAJ. Spartan Books. 648 pp.(cit. on p. 29)

Rošić, Silvana, Rachel Amouroux, Cristina E. Requena, et al. (Mar. 2018). “Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity”. In: *Nature Genetics* 50.3, pp. 452–459. DOI: 10.1038_s41588 * 018 * 0061 * 8. (cit. on p. 17)

Ross, Paul D., Jacky Guy, Jim Selfridge, et al. (Aug. 9, 2016). "Exclusive expression of MeCP2 in the nervous system distinguishes between brain and peripheral Rett syndrome-like phenotypes". In: *Human Molecular Genetics*, ddw269. DOI: 10.1093_hmg_ddw269. (cit. on p. 36, 43)

Roth, Sharon Y., John M. Denu, and C. David Allis (June 2001). "Histone Acetyltransferases". In: *Annual Review of Biochemistry* 70.1, pp. 81–120. DOI: 10.1146_annurev.biochem.70.1.81. (cit. on p. 14)

Rouet, Philippe, Fatima Smih, and Maria Jasin (Dec. 1, 1994). "Introduction of Double-Strand Breaks into the Genome of Mouse Cells by Expression of a Rare-Cutting Endonuclease". In: *Molecular and Cellular Biology* 14.12, pp. 8096–8106. DOI: 10.1128_mcb.14.12.8096 * 8106.1994. (cit. on p. 19)

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (Oct. 1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038_323533a0. (cit. on p. 29)

Sakaki-Yumoto, Masayo, Chiyoko Kobayashi, Akira Sato, et al. (Aug. 1, 2006). "The murine homolog of *SALL4* , a causative gene in Okihiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with *Sall1* in anorectal, heart, brain and kidney development". In: *Development* 133.15, pp. 3005–3013. DOI: 10.1242_dev.02457. (cit. on p. 38)

Sanger, F. and A.R. Coulson (May 1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *Journal of Molecular Biology* 94.3, pp. 441–448. DOI: 10.1016_0022 * 2836.75/90213 * 2. (cit. on p. 20)

Schloss, Jeffery A (Oct. 2008). "How to get genomes at one ten-thousandth the cost". In: *Nature Biotechnology* 26.10, pp. 1113–1115. DOI: 10.1038_nbt1008 * 1113. (cit. on p. 20)

Schreiber, Jacob, Timothy Durham, Jeffrey Bilmes, et al. (Mar. 30, 2020). "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome". In: *Genome Biology* 21.1, p. 81. DOI: 10.1186_s13059*020*01977*6. (cit. on p. 39)

Schuettengruber, Bernd and Giacomo Cavalli (Nov. 1, 2009). "Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice". In: *Development* 136.21, pp. 3531–3542. DOI: 10.1242_dev.033902. (cit. on p. 11, 12)

Schwalb, Björn, Margaux Michel, Benedikt Zacher, et al. (June 3, 2016). "TT-seq maps the human transient transcriptome". In: *Science* 352.6290, pp. 1225–1228. DOI: 10.1126_science.aad9841. (cit. on p. 22)

Shah, Rohan N., Adrian T. Grzybowski, Jimmy Elias, et al. (Sept. 10, 2021). *Re-evaluating the role of nucleosomal bivalency in early development*. Pages: 2021.09.09.458948 Section: New Results. DOI: 10.1101_2021.09.09.458948. (cit. on p. 40)

Shapley, L. S. (Dec. 31, 1953). "17. A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton University Press, pp. 307–318. DOI: 10.1515_9781400881970 * 018. (cit. on p. 31)

Shilatifard, Ali (July 7, 2012). "The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis". In: *Annual*

Review of Biochemistry 81.1, pp. 65–95. DOI: 10.1146_annurev * biochem * 051710 * 134100. (cit. on p. 13)

Shin Voo, Kui, Diana L. Carlone, Britta M. Jacobsen, et al. (Mar. 1, 2000). "Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1". In: *Molecular and Cellular Biology* 20.6, pp. 2108–2121. DOI: 10.1128_MCB.20.6.2108 * 2121.2000. (cit. on p. 37)

Shogren-Knaak, Michael, Haruhiko Ishii, Jian-Min Sun, et al. (Feb. 10, 2006). "Histone H4-K16 Acetylation Controls Chromatin Structure and Protein Interactions". In: *Science* 311.5762, pp. 844–847. DOI: 10.1126_science.1124000. (cit. on p. 12)

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (Oct. 12, 2019). *Learning Important Features Through Propagating Activation Differences*. DOI: 10.48550_arXiv.1704.02685. arXiv: 1704.02685[cs]. (cit. on p. 33)

Simon, Jeffrey A. and Robert E. Kingston (Oct. 2009). "Mechanisms of Polycomb gene silencing: knowns and unknowns". In: *Nature Reviews Molecular Cell Biology* 10.10, pp. 697–708. DOI: 10.1038_nrm2763. (cit. on p. 11)

Skene, Peter J and Steven Henikoff (Jan. 16, 2017). "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites". In: *eLife* 6, e21856. DOI: 10.7554_eLife.21856. (cit. on p. 22)

Skene, Peter J., Robert S. Illingworth, Shaun Webb, et al. (Feb. 2010). "Neuronal MeCP2 Is Expressed at Near Histone-Octamer Levels and Globally Alters the Chromatin State". In: *Molecular Cell* 37.4, pp. 457–468. DOI: 10.1016_j.molcel.2010.01.030. (cit. on p. 36)

Smith, Temple F. (Apr. 1990). "The history of the genetic sequence databases". In: *Genomics* 6.4, pp. 701–707. DOI: 10.1016_0888 * 7543.90/90509 * S. (cit. on p. 24)

Sneppen, Kim and Leonie Ringrose (May 13, 2019). "Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent". In: *Nature Communications* 10.1. Publisher: Nature Publishing Group, p. 2133. DOI: 10.1038_s41467 * 019 * 10130 * 2. (cit. on p. 40)

Soto, Luis F., Zhaorong Li, Clarissa S. Santoso, et al. (Feb. 2022). "Compendium of human transcription factor effector domains". In: *Molecular Cell* 82.3, pp. 514–526. DOI: 10.1016_j.molcel.2021.11.007. (cit. on p. 8, 9)

Sperlazza, M. Jeannette, Stephanie M. Bilinovich, Leander M. Sinanan, et al. (May 2017). "Structural Basis of MeCP2 Distribution on Non-CpG Methylated and Hydroxymethylated DNA". In: *Journal of Molecular Biology* 429.10, pp. 1581–1594. DOI: 10.1016_j.jmb.2017.04.009. (cit. on p. 35)

Stadler, Michael B., Rabih Murr, Lukas Burger, et al. (Dec. 22, 2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions". In: *Nature* 480.7378, pp. 490–495. DOI: 10.1038_nature10716. (cit. on p. 17)

Stark, Rory, Marta Grzelak, and James Hadfield (Nov. 2019). "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11, pp. 631–656. DOI: 10.1038_s41576 * 019 * 0150 * 2. (cit. on p. 21)

Stein, Chad B., Andrew R. Field, Claudia A. Mimoso, et al. (Nov. 2022). "Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci". In: *Molecular Cell* 82.22, 4232–4245.e11. DOI: 10.1016_j.molcel.2022.10.004. (cit. on p. 40, 41)

Sternberg, Samuel H., Sy Redding, Martin Jinek, et al. (Mar. 2014). "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9". In: *Nature* 507.7490, pp. 62–67. DOI: 10.1038_nature13011. (cit. on p. 19)

Stigler, Stephen M. (May 1, 1981). "Gauss and the Invention of Least Squares". In: *The Annals of Statistics* 9.3. DOI: 10.1214_aos_1176345451. (cit. on p. 26)

Tate, Courtney M., Jeong-Heon Lee, and David G. Skalnik (Jan. 2010). "CXXC finger protein 1 restricts the Setd1A histone H3K4 methyltransferase complex to euchromatin". In: *The FEBS Journal* 277.1, pp. 210–223. DOI: 10.1111_j.1742 * 4658.2009.07475.x. (cit. on p. 37)

Teytelman, Leonid, Deborah M. Thurtle, Jasper Rine, et al. (Nov. 12, 2013). "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins". In: *Proceedings of the National Academy of Sciences* 110.46, pp. 18602–18607. DOI: 10.1073_pnas.1316064110. (cit. on p. 22)

Thakur, Jitendra, Jenika Packiaraj, and Steven Henikoff (Apr. 21, 2021). "Sequence, Chromatin and Evolution of Satellite DNA". In: *International Journal of Molecular Sciences* 22.9, p. 4309. DOI: 10.3390_ijms22094309. (cit. on p. 36, 58)

The ENCODE Project Consortium (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038_nature11247. (cit. on p. 1, 14)

The ENCODE Project Consortium, Federico Abascal, Reyes Acosta, et al. (July 30, 2020). "Perspectives on ENCODE". In: *Nature* 583.7818, pp. 693–698. DOI: 10.1038_s41586 * 020 * 2449 * 8. (cit. on p. 2)

Thomson, John P., Peter J. Skene, Jim Selfridge, et al. (Apr. 2010). "CpG islands influence chromatin structure via the CpG-binding protein Cfp1". In: *Nature* 464.7291, pp. 1082–1086. DOI: 10.1038_nature08924. (cit. on p. 37)

Tillotson, Rebekah and Adrian Bird (Mar. 2020). "The Molecular Basis of MeCP2 Function in the Brain". In: *Journal of Molecular Biology* 432.6, pp. 1602–1623. DOI: 10.1016_j.jmb.2019.10.004. (cit. on p. 35)

Tillotson, Rebekah, Justyna Cholewa-Waclaw, Kashyap Chhatbar, et al. (Mar. 2021). "Neuronal non-CG methylation is an essential target for MeCP2 function". In: *Molecular Cell* 81.6, 1260–1275.e12. DOI: 10.1016_j.molcel.2021.01.011. (cit. on p. 36, 58)

Tong, Jeffrey K., Christian A. Hassig, Gavin R. Schnitzler, et al. (Oct. 1998). "Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex". In: *Nature* 395.6705, pp. 917–921. DOI: 10.1038_27699. (cit. on p. 15)

Uptain, S. M., C. M. Kane, and M. J. Chamberlin (June 1997). "BASIC MECHANISMS OF TRANSCRIPT ELONGATION AND ITS REGULATION". In: *Annual Review of Biochemistry* 66.1, pp. 117–172. DOI: 10.1146_annurev.biochem.66.1.117.(cit. on p. 6)

Vakoc, Christopher R., Mira M. Sachdeva, Hongxin Wang, et al. (Dec. 1, 2006). "Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin". In: *Molecular and Cellular Biology* 26.24, pp. 9185–9195. DOI: 10.1128_MCB.01529 * 06. (cit. on p. 11)

Van Esch, Hilde, Marijke Bauters, Jaakko Ignatius, et al. (Sept. 2005). "Duplication of the MECP2 Region Is a Frequent Cause of Severe Mental Retardation and Progressive Neurological Symptoms in Males". In: *The American Journal of Human Genetics* 77.3, pp. 442–453. DOI: 10.1086_444549. (cit. on p. 18)

Vermeulen, Michiel, Klaas W. Mulder, Sergei Denissov, et al. (Oct. 2007). "Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4". In: *Cell* 131.1, pp. 58–69. DOI: 10.1016_j.cell.2007.08.016. (cit. on p. 11)

Vitaliano-Prunier, Adeline, Alexandra Menant, Maria Hobeika, et al. (Nov. 2008). "Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation". In: *Nature Cell Biology* 10.11, pp. 1365–1371. DOI: 10.1038_ncb1796. (cit. on p. 12)

Voigt, Philipp, Wee-Wei Tee, and Danny Reinberg (June 15, 2013). "A double take on bivalent promoters". In: *Genes & Development* 27.12, pp. 1318–1338. DOI: 10.1101_gad.219626.113. (cit. on p. 40)

Wade, Paul A., Anne Gegonne, Peter L. Jones, et al. (Sept. 1999). "Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation". In: *Nature Genetics* 23.1, pp. 62–66. DOI: 10.1038_12664. (cit. on p. 15)

Wade, Paul A., Peter L. Jones, Danielle Vermaak, et al. (July 1998). "A multiple subunit Mi-2 histone deacetylase from Xenopus laevis cofractionates with an associated Snf2 superfamily ATPase". In: *Current Biology* 8.14, pp. 843–848. DOI: 10.1016_S0960 * 9822.98/70328 * 8. (cit. on p. 15)

Wagner, Eric J., Liang Tong, and Karen Adelman (Feb. 2023). "Integrator is a global promoter-proximal termination complex". In: *Molecular Cell* 83.3, pp. 416–427. DOI: 10.1016_j.molcel.2022.11.012. (cit. on p. 13, 40)

Wang, Hua, Zheng Fan, Pavel V. Shliaha, et al. (Mar. 2023). "H3K4me3 regulates RNA polymerase II promoter-proximal pause-release". In: *Nature* 615.7951. Publisher: Nature Publishing Group, pp. 339–348. DOI: 10.1038_s41586 * 023 * 05780 * 8.(cit. on p. 19, 40, 41)

Wang, Zhibin, Chongzhi Zang, Kairong Cui, et al. (Sept. 2009). "Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes". In: *Cell* 138.5, pp. 1019–1031. DOI: 10.1016_j.cell.2009.06.049. (cit. on p. 12, 14)

Watson, J. D. and F. H. C. Crick (Apr. 25, 1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356, pp. 737–738. DOI: 10.1038_171737a0. (cit. on p. 5)

Weber, Michael, Ines Hellmann, Michael B Stadler, et al. (Apr. 2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome". In: *Nature Genetics* 39.4, pp. 457–466. DOI: 10.1038_ng1990.(cit. on p. 17)

Weinberg, Daniel N., Simon Papillon-Cavanagh, Haifen Chen, et al. (Sept. 12, 2019). "The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape". In: *Nature* 573.7773, pp. 281–286. DOI: 10.1038_s41586 * 019 * 1534 * 3. (cit. on p. 11, 18)

Wheeler, H (1904). "Researches on pyrimidine derivatives". In: *J. Am. Chem. Soc* 31, pp. 591–606. (cit. on p. 16)

Wiedenheft, Blake, Samuel H. Sternberg, and Jennifer A. Doudna (Feb. 2012). "RNA-guided genetic silencing systems in bacteria and archaea". In: *Nature* 482.7385, pp. 331–338. DOI: 10.1038_nature10886. (cit. on p. 19)

Wilkins, M. H. F., A. R. Stokes, and H. R. Wilson (Apr. 25, 1953). "Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids". In: *Nature* 171.4356, pp. 738–740. DOI: 10.1038_171738a0. (cit. on p. 5)

Wong, Justin J. -L., Dadi Gao, Trung V. Nguyen, et al. (May 8, 2017). "Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment". In: *Nature Communications* 8.1, p. 15134. DOI: 10.1038_ncomms15134. (cit. on p. 37, 43)

Wong, Justin J.-L., William Ritchie, Olivia A. Ebner, et al. (Aug. 2013). "Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation". In: *Cell* 154.3, pp. 583–595. DOI: 10.1016_j.cell.2013.06.052. (cit. on p. 43)

Wu, Xudong, Jens Vilstrup Johansen, and Kristian Helin (Mar. 2013). "Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation". In: *Molecular Cell* 49.6, pp. 1134–1146. DOI: 10.1016_j.molcel.2013.01.016. (cit. on p. 38)

Wyatt, G. R. (Aug. 1950). "Occurrence of 5-Methyl-Cytosine in Nucleic Acids". In: *Nature* 166.4214, pp. 237–238. DOI: 10.1038_166237b0. (cit. on p. 16)

— (May 1, 1951). "Recognition and estimation of 5-methylcytosine in nucleic acids". In: *Biochemical Journal* 48.5, pp. 581–584. DOI: 10.1042_bj0480581. (cit. on p. 16)

Xu, Lan, Christopher K Glass, and Michael G Rosenfeld (Apr. 1999). "Coactivator and corepressor complexes in nuclear receptor function". In: *Current Opinion in Genetics & Development* 9.2, pp. 140–147. DOI: 10.1016_S0959 * 437X.99/80021 * 5. (cit. on p. 14)

Xue, Yutong, Jiemin Wong, G.Tony Moreno, et al. (Dec. 1998). "NURD, a Novel Complex with Both ATP-Dependent Chromatin-Remodeling and Histone Deacetylase Activities". In: *Molecular Cell* 2.6, pp. 851–861. DOI: 10.1016_S1097 * 2765.00/80299 * 3. (cit. on p. 15)

Yamashita, Kazunari, Akira Sato, Makoto Asashima, et al. (Feb. 2007). "Mouse homolog of SALL1, a causative gene for Townes–Brocks syndrome, binds to A/T-rich sequences in pericentric heterochromatin via its C-terminal zinc finger domains". In: *Genes to Cells* 12.2, pp. 171–182. DOI: 10.1111_j.1365 * 2443.2007.01042.x. (cit. on p. 38)

Young, Juan I., Eugene P. Hong, John C. Castle, et al. (Dec. 6, 2005). "Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2". In: *Proceedings of the National Academy of Sciences* 102.49, pp. 17551–17558. DOI: 10.1073_pnas.0507856102. (cit. on p. 35, 37)

Zemach, Assaf and Daniel Zilberman (Sept. 2010). "Evolution of Eukaryotic DNA Methylation and the Pursuit of Safer Sex". In: *Current Biology* 20.17, R780–R785. DOI: 10.1016_j.cub.2010.07.007. (cit. on p. 16)

Zhang, Y., H.-H. Ng, H. Erdjument-Bromage, et al. (Aug. 1, 1999). "Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation". In: *Genes & Development* 13.15, pp. 1924–1935. DOI: 10.1101_gad.13.15.1924. (cit. on p. 15)

Zhang, Yi, Gary LeRoy, Hans-Peter Seelig, et al. (Oct. 1998). "The Dermatomyositis-Specific Autoantigen Mi2 Is a Component of a Complex Containing Histone Deacetylase and Nucleosome Remodeling Activities". In: *Cell* 95.2, pp. 279–289. DOI: 10.1016_S0092 * 8674.00/81758 * 4. (cit. on p. 15)

Zheng, Hai, Yilun Qi, Shibin Hu, et al. (Nov. 27, 2020). "Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase". In: *Science* 370.6520. Publisher: American Association for the Advancement of Science, eabb5872. DOI: 10.1126_science.abb5872. (cit. on p. 13)