Francis Crick Institute &
Institute of Clinical Sciences, Imperial College London

Investigation of the gene regulatory logic encoded in genome sequences using machine learning

Sebastian Steinhauser

February 2020

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London

“Da steh ich nun, ich armer Tor, und bin so klug als wie zuvor.”
- Johann Wolfgang von Goethe, Faust (Part One)

DECLARATION

I, Sebastian Steinhauser, hereby confirm that the submitted PhD thesis is my own work. Each results chapter contains a declaration of my work and acknowledges contributions from collaborators. To the best of my knowledge, all information and figures from published works are appropriately referenced.

No parts of this PhD thesis were previously submitted for any other degree or qualification at any university or institution.

This PhD thesis complies with the Imperial College Thesis guidelines. The Abstract does not exceed 300 words and the entire thesis does not exceed 100,000 words.

Sebastian Steinhauser
February 2020

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

ACKNOWLEDGEMENTS

Firstly, I want to thank my PhD supervisor Nicholas Luscombe for all the opportunities during my PhD in his lab, his feedback and for supporting me and my projects through the highs and lows. Secondly, I also want to thank my university supervisor Boris Lenhard for his support, feedback, letting me attend his weekly lab meetings and various other interesting discussions. Thirdly, I would like to thank James Briscoe for the opportunity to work on exciting projects originating from his lab, his support, valuable insights and discussions. Fourthly, I want to say thank you to all the members of my thesis committee (Boris, James and Jernej Ule) who sacrificed their valuable time to listen to my ongoing research and provide helpful feedback. Finally, I want to say thank you to Jorge Ferrer and Martin Hemberg for agreeing to the last minute viva before the Coronavirus lockdown and further for the nice discussion and valuable feedback.

Someone who deserves tremendous credit is Aylin Cakiroglu: thank you so much for supervising me; all the many things I learned from you about machine learning; our prison walks and most of all, for your friendship. I want to thank Vicki Metzis: the best experimental partner and collaborator a bioinformatician could wish for. Your enthusiasm and passion for developmental biology are infectious and I learned a lot about it from you. Yet another person who contributed to this thesis was Mahmoud-Reza Rafiee and I would like to thank him for generating the ChIP and STARR-seq data used in Chapter 2 and 3.

I also want to say thank you to Charlotte, Daniel F., Fede, George, Nobby, Igor and Miha for everything: our scientific and non-scientific discussions; our Friday Club adventures; the great evenings, meals and drinks we shared and your friendship - it does mean a lot to me. Further I want to thank the many members from the Briscoe, Luscombe and Ule Lab with whom I crossed paths or had a drink with at the Crick bar. I also want to thank people from my second Lab including Nevena, Fabio, Radina, Dunja, Leonie and Alex for valuable feedback on my research, the discussions we had and their friendship. The LMS people, including the Lenhard Lab, I want to thank for some awesome Cost Cutter sessions and other nice outings.

A final kudos to everyone who sacrificed their time to read this thesis or anything else I've written including: Aylin, Boris, Charlotte, James, Nick, Nobby and Vicki.

Besonderer Dank gebührt auch meinen zwei besten Freunden auf dieser Welt, Nils und Ruben. Euch beiden danke für all die Diskussionen, die Hilfe in allen Lebenslagen und auch das ein oder andere gemeinsame Bier. Außerdem möchte ich dem ganzen MoBi Alumni Team, Georg, Jonathan, Felix, Fabian und Roman danken für ihre Freundschaft, die schönen Wochenenden jedes Jahr in Heidelberg und wann man sich sonst noch zum Gedankenaustausch oder auf ein Bierchen trifft.

Außerdem möchte ich meinen Freunden aus der Heimat danken, besonders Mario, Uwe und Pascal, die immer ein offenes Ohr für mich und meine Doktorarbeit hatten. Den dreien und dem ganzen DDSV für viele urige und schöne Abende mit dem ein oder anderen Bier.

An dieser Stelle möchte ich mich besonders bei meinem Vater und meiner Mutter für alles bedanken: die finanzielle, ideelle und sonstige Hilfe in allen Lebenslagen über den Verlauf meines ganzen Studiums. Vor allem dir Mama möchte ich noch einmal besonders danken für alles, was du für mich getan hast. Ich wäre niemals soweit gekommen ohne deine Unterstützung. Ich möchte außerdem meinem Opa Franz für alles danken, das er in meinem Leben für mich getan hat, und mir dabei geholfen hat, diesen Punkt zu erreichen. Ich wünschte mir so sehr, dass du das noch hättest miterleben können. Außerdem möchte ich mich bei Norbert bedanken für all die Hilfe während meines Studiums, die vielen nächtlichen Fahrten zum Flughafen und den ein oder anderen schönen Besen Abend. Zum Schluss möchte ich dem Rest meiner Familie danken für alles sowohl Materielle als auch Ideelle, welches zu meinem Studium vom Bachelor bis zur Doktorarbeit beigetragen hat.

Am allerwichtigsten: Ich möchte mich bei der Frau meines Lebens, Verena, bedanken. Ohne deine kontinuierliche Unterstützung und deinen Zuspruch hätte ich diese Arbeit nicht geschafft. Trotz allem Stress und Gemotze waren es am Ende gerade wegen dir und unseren gemeinsamen Erlebnissen 4 schöne Jahre, die ich niemals missen wollte.

ABSTRACT

Spatiotemporal regulation of gene expression in multicellular organisms is mediated by cell-type specific regulatory elements called transcriptional enhancers. Their activity is regulated by the coordinated binding of transcription factors (TFs) to specific DNA sequences encoded within these elements. In this PhD thesis, we investigate transcriptional enhancers in the context of pluripotency and early neural development.

In Chapter 2, we investigate how enhancer activity is encoded in the DNA sequence underlying TF binding sites. To this end, enhancer activities of NANOG binding sites in naive mouse embryonic stem cells (mESCs) were measured using ChIP STARR-seq. This revealed two distinct classes of NANOG binding sites, active enhancers and inactive binding sites. Predictive modelling of these classes uncovered sequence features which account for these differences such as the ESRRB motif.

In Chapter 3, we examined differential enhancer activities of NANOG binding sites between naive and primed mESC using ChIP STARR-seq. Comprehensive analysis of various chromatin genomic features revealed chromatin accessibility changes as a major predictor of differential enhancer activity. Predictive modelling using only DNA sequence features uncovered potential transcriptional regulators that define enhancer activity in the two cellular states, such as TCF3, TCF4 and ESRRB for naive mESC and ZIC3 for primed cells.

In Chapter 4, we studied the timing of neural induction during early nervous system development along the anterior-posterior axis using information about enhancer accessibilities. Changes in open-chromatin states during neural development were measured by applying ATAC-seq to an *in vitro* model system. Analysis of resulting data revealed that for posterior spinal cord induction, cells initially commit to a posterior regional identity before acquiring neural identity. Moreover, our analysis highlighted the role of CDX2 during this process which represses hindbrain and activates spinal cord related transcriptional programs. Together, this Chapter provides conclusively evidence for the dual origin of neural progenitor cells.

TABLE OF CONTENTS

List of figures **xvii**

List of tables **xxi**

Acronyms **xxvi**

1 Introduction **1**

1.1 Mechanisms of transcriptional regulation 1

1.1.1 Transcription of genes to messenger RNA (mRNA) 1

1.1.2 Regulatory elements encode information to regulate gene expression 2

1.1.3 Transcription factors and co-factors 7

1.1.4 Chromatin structure and Histone PTMs 8

1.1.5 Role of DNA modifications in transcriptional regulation 9

1.1.6 3D genome architecture and promoter-enhancer interactions 10

1.2 Sequencing methods in regulatory genomics and their computational analysis 12

1.2.1 Profiling of TF binding and Histone modifications 12

1.2.2 Chromatin accessibility profiling 14

1.2.3 Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) 14

1.2.4 Methods for DNA methylation profiling 15

1.2.5 Detection of Chromatin-Chromatin interactions 16

1.2.6 Measuring gene expression using RNA sequencing (RNA-seq) . . . 17

1.3 Profiling nascent transcription 18

1.3.1 Comparison of different sequencing data based enhancer definitions 19

1.4 Outline and aim of the thesis 22

2 Deciphering the sequence encoded regulatory logic of NANOG binding enhancers **23**

2.1 Introduction . . . 25
2.2 Methods . . . 27
2.2.1 Experimental procedures . . . 27
2.2.2 ChIP-seq pre-processing & analysis . . . 27
2.2.3 ChIP STARR-seq pre-processing & analysis . . . 28
2.2.4 CUT&RUN pre-processing & analysis . . . 29
2.2.5 ChIP-exo pre-processing & analysis . . . 29
2.2.6 RNA-seq pre-processing & analysis . . . 30
2.2.7 GRO-seq pre-processing & analysis . . . 30
2.2.8 ATAC-seq pre-processing & analysis . . . 30
2.2.9 Genome browser track plots . . . 31
2.2.10 Capture Hi-C analysis . . . 31
2.2.11 Peak-gene annotation and enrichment analysis . . . 32
2.2.12 Core-promoter element enrichment . . . 32
2.2.13 Binary classification of active enhancers and inactive TFBS . . . 32
2.3 Results . . . 37
2.3.1 Defining high-confidence NANOG TFBS from ChIP-seq and ChIP STARR-seq . . . 37
2.3.2 STARR-seq quality control analysis . . . 40
2.3.3 Defining *bona fide* NANOG binding enhancers by comparison of vector library and STARR-seq . . . 45
2.3.4 Transcription factor binding at active enhancers and inactive TFBS . . . 48
2.3.5 Chromatin features of active enhancers and inactive TFBS . . . 53
2.3.6 Nascent transcription of ncRNA from active enhancers and inactive TFBS . . . 57
2.3.7 Chromatin-chromatin interactions of active enhancers and inactive TFBS . . . 58
2.3.8 Influence of active enhancers and inactive TFBS on gene expression . . . 61
2.3.9 Differences in core-promoter architecture of active enhancers and inactive TFBS . . . 63
2.3.10 Predicting active enhancers and inactive TFBS from DNA sequence alone . . . 64

2.3.11 Knockout of *Esrrb* leads to loss of H3K27ac at active Nanog binding enhancers . . . 76
2.4 Discussion . . . 79

3 Assessing condition-specific enhancer activity of NANOG binding sites 85
3.1 Introduction . . . 87
3.2 Methods . . . 90
3.2.1 Experimental procedures . . . 90
3.2.2 Regulatory genomics sequencing data preprocessing and analysis . . . 90
3.2.3 Definition of 2i- and serum-specific enhancer elements . . . 90
3.2.4 DNA methylation analysis . . . 91
3.2.5 Proteomics analysis . . . 91
3.2.6 Prediction of condition-specific enhancers using chromatin features . . . 91
3.2.7 Prediction of condition-specific enhancers activity using sequence features . . . 92
3.3 Results . . . 94
3.3.1 Definition of NANOG binding enhancers specific to 2i and serum ESC conditions . . . 94
3.3.2 Binding of pluripotency TFs at naive (2i) and primed (serum) ESC enhancers . . . 99
3.3.3 Differences in the epigenetic landscape around naive (2i) and primed (serum) ESC enhancers . . . 101
3.3.4 Open chromatin dynamics of naive (2i) and primed (serum) ESC enhancers . . . 104
3.3.5 DNA methylation dynamics at (2i) and primed (serum) ESC enhancers . . . 105
3.3.6 Nascent eRNA transcription at (2i) and primed (serum) ESC enhancers . . . 107
3.3.7 Repetitive element overlap with naive (2i) and primed (serum) ESC enhancers . . . 108
3.3.8 Chromatin-chromatin interactions of naive (2i) and primed (serum) ESC enhancers . . . 109
3.3.9 Gene expression of naive (2i) and primed (serum) enhancers target genes . . . 111

3.3.10 Prediction of condition-specific enhancer activity using chromatin features . . . 112
3.3.11 Prediction of condition-specific enhancer activity using sequence features . . . 115
3.4 Discussion . . . 125

4 Chromatin accessibility dynamics during early neurogenesis reveal timing of axis allocation and neural induction 129
4.1 Introduction . . . 131
4.2 Methods . . . 135
4.2.1 ATAC-seq and ChIP-seq . . . 135
4.2.2 ATAC-seq analysis . . . 135
4.2.3 ChIP-seq analysis . . . 138
4.2.4 RNA-seq analysis . . . 139
4.2.5 Microarray analysis . . . 139
4.2.6 Chromatin-Chromatin interaction analysis . . . 139
4.3 Results . . . 140
4.3.1 *In vitro* model of neural progenitor development generates cells with different AP-axis identities . . . 140
4.3.2 ATAC-seq quality control analysis . . . 142
4.3.3 Differential chromatin accessibility defines neural cell fates . . . 145
4.3.4 Regions of dynamic accessibility correspond to enhancer elements . . . 148
4.3.5 *In vitro* enhancer signatures exhibit similar accessibility dynamics *in vivo* . . . 149
4.3.6 Motif and ChIP-seq enrichment analysis uncovered major regulators of observable chromatin dynamics . . . 151
4.3.7 Chromatin dynamics induced by WNT depends on developmental time . . . 154
4.3.8 NMPs show transient chromatin accessibility induced by WNT signalling . . . 156
4.3.9 CDX2 is the master regulator of NMP specific regulatory elements . . . 158
4.3.10 CDX2 inhibits hindbrain in favour of spinal cord identity . . . 161
4.3.11 CDX2, not WNT signalling, provides cells with the competence to generate spinal cord . . . 164
4.4 Discussion . . . 168

5 Conclusions and future directions **173**
5.1 Improving ChIP STARR-seq to achieve high resolution readouts . . . 174
5.2 Validation of ChIP STARR-seq and predicted regulators of enhancer activity . . . 175
5.3 Definition of *bona fide* transcriptional enhancers from ATAC-seq . . . 176
5.4 Resolving cellular heterogeneity using single cell ATAC-seq . . . 177
5.5 Concluding remarks . . . 178

Appendix A Supplementary Figures **217**
A.1 Chapter2 Supplementary Figures . . . 218
A.2 Chapter3 Supplementary Figures . . . 221

Appendix B Supplementary Tables **223**
B.1 Chapter 2 and 3 Supplementary Tables . . . 224
B.2 Chapter 4 Supplementary Tables . . . 228

LIST OF FIGURES

1.1 Overview of sequencing methods used to define transcriptional enhancers. . 20

2.1 ChIP STARR-seq experiment to determine active enhancers from NANOG binding sites. 26
2.2 Definition of NANOG consensus binding sites from ChIP-seq and STARR-seq vector library. 38
2.3 NANOG motif enrichment for putative NANOG enhancer set. 40
2.4 Quality control analysis of vector library and STARR-seq samples. 42
2.5 Comparative analysis between ChIP-seq, STARR-seq vector library and STARR-seq samples. 44
2.6 Differential analysis between vector library and STARR-seq to define active enhancers. 46
2.7 Examples for an active NANOG binding enhancer and inactive NANOG binding site. 47
2.8 NANOG, OCT4 and SOX2 binding, measured by ChIP-seq, at active enhancers and inactive TFBS. 49
2.9 CUT&RUN signal for NANOG, OCT4 and SOX2 at active enhancers and inactive TFBS. 50
2.10 TF and nucleosome occupancy as measured by ATAC-seq in active enhancers and inactive binding sites. 51
2.11 Differential enrichment analysis of TFs between active enhancers and inactive binding sites. 53
2.12 Histone PTM ChIP-seq and ATAC-seq signal around active enhancers and inactive TFBS. 54
2.13 Differential enrichment analysis of histone modifications between active enhancers and inactive TFBS. 55
2.14 Differential enrichment analysis of chromatin associated proteins between active enhancers and inactive TFBS. 57
2.15 Nascent transcription of ncRNAs at active enhancers and inactive binding sites. 58

2.16 Chromatin-chromatin interactions of active enhancers and inactive binding sites. 60
2.17 Expression of active enhancers and inactive TFBS target genes. 62
2.18 Enrichment of core-promoter elements in target gene sets assigned to active enhancer and inactive TFBS. 64
2.19 Proteomics meta-analysis of published NANOG co-binding proteins and ESC Chromatome. 66
2.20 Predictive power of single PWMs for enhancer activity of NANOG binding sites. 67
2.21 Performance benchmark on held-out test set of classification models using different metrics. 70
2.22 Feature importance ranking for GBM trained on NANOG co-binding TFs. . 71
2.23 ESRRB and TBX3 binding at active enhancers and inactive TFBs. 72
2.24 DeepLIFT importance scores relative to ChIP-exo summits. 73
2.25 TF-MoDisco de novo discovered TF motifs from DeepLIFT importance scores. 74
2.26 Feature importance of TF-MoDisco de novo discovered TF motifs. 75
2.27 Binding of ESRRB to TF-MoDisco de novo discovered motifs. 76
2.28 H3K27ac ChIP-seq from ESC with inducible knockouts of either *Nr5a2* or *Esrrb*. 78

3.1 ChIP STARR-seq experiment to determine condition-specific *bona fide* transcriptional enhancers from NANOG binding sites. 88
3.2 Definition of naive (2i) and primed (serum) NANOG binding pluripotency enhancers. 95
3.3 Example for naive (2i) mESC specific transcriptional enhancer at the *Prdm14* locus. 97
3.4 Example for primed (serum) mESC specific transcriptional enhancer at the *Arl4d* locus. 98
3.5 NANOG binding to naive (2i) and primed (serum) enhancer. 99
3.6 NANOG, OCT4 and SOX2 binding to naive (2i) and primed (serum) enhancers. 100
3.7 Epigenetic landscape of naive (2i) and primed (serum) enhancers. 103
3.8 Open chromatin dynamics of naive (2i) and primed (serum) enhancers. . . . 105
3.9 Methylation levels at naive (2i) and primed (serum) enhancers. 106
3.10 Nascent eRNA transcription at naive (2i) and primed (serum) enhancers. . . 108
3.11 Enrichment of sequence repeats in naive (2i) and primed (serum) enhancer sets. 109

3.12 Chromatin-chromatin interactions of naive (2i) and primed (serum) enhancers. 110
3.13 Enhancer target gene expression from 2i/serum RNA-seq and pathway enrichment. 112
3.14 Identification of chromatin associated proteins predictive for differential enhancer activity. 113
3.15 Predictive power of chromatin features for differential enhancer activity. . . 114
3.16 Prioritisation of potential TFs involved in differential enhancer activity. . . . 117
3.17 Performance benchmark on hold-out test set of classification models using different metrics. 120
3.18 Feature importance ranking for GBM trained on all proteomics TFs. 122
3.19 TCF7L1, ZIC3 and ESRRB binding, as measured by ChIP-seq, at 2i- and serum-specific enhancers. 124

4.1 Proposed hypothesis for neural induction and AP-axis pattering during early neurogenesis. 133
4.2 *In vitro* differentiation of mESC to NPC with different regional identities along the AP axis. 140
4.3 Examples enhancers displaying changes in chromatin accessibilities during ESC to NPC differentiation. 141
4.4 Quality control analysis for all in-house generated ATAC-seq samples. . . . 144
4.5 Genome-wide differential analysis of ATAC-seq accessibility uncovers dynamic sites. 146
4.6 Clustering of differential accessible sites using a self-organising map uncovered chromatin dynamics. 147
4.7 Examples for enhancer elements specific for NPC AP-axis regional identity. 148
4.8 Overlap of neural and AP regional identity clusters with genome annotation. 149
4.9 Comparison of ATAC-seq derived from *in vitro* and *in vivo* NPCs. 151
4.10 Enrichment of TF motifs and ChIP-seq peaks in regulatory elements defining AP-axis regional identity. 153
4.11 Gene expression profiling on NPC with delayed WNT induction. 155
4.12 ATAC-seq analysis of day 5 hindbrain-plus NPCs. 156
4.13 NMP enhancers are enriched for CDX2 binding sites. 158
4.14 NMP enhancer accessibility is not regulated by T/BRA. 159
4.15 CDX2 regulates NMP enhancer accessibility and spinal cord identity. 160

4.16 *Cdx* KO cells differentiated towards hindbrain rather than posterior spinal cord identity. 161
4.17 CDX2 binding sites overlap with H-SC and hindbrain enhancers. 162
4.18 CDX2 binds in different chromatin contexts to induce spinal cord and repress hindbrain identity. 164
4.19 CDX2 can replace WNT signalling to induce spinal cord and repress hindbrain identity. 166
4.20 Proposed model of axial allocation and neural induction during early neurogenesis. 170

A.1 Grid-search results from GBM hyperparameter optimisation. 218
A.2 Grid-search results for gkm-SVM hyperparameter optimisation. 219
A.3 Model training history for CNN with best architecture from hyperparameter optimisation. 220
A.4 Grid-search results from GBM hyperparameter optimisation. 221
A.5 Grid-search results for gkm-SVM hyperparameter optimisation. 222

LIST OF TABLES

1.1 List of relevant histone PTMs and their associated functions. 9

2.1 GBM hyperparameter search space. 33
2.2 gkm-SVM hyperparameter search space. 34
2.3 CNN hyperparameter search space. 35

3.1 Summary of sequencing methods, analysis methods and sample lists. 90

B.1 ChIP STARR-seq datasets used in Chapter 2 & 3. 224
B.2 ChIP-exo datasets used in Chapter 2 & 3. 224
B.3 CUT&RUN datasets used in Chapter 2. 224
B.4 ChIP-seq datasets used in Chapter 2 & 3. 225
B.5 ATAC-seq datasets used in Chapter 2 & 3. 226
B.6 GRO-seq datasets used in Chapter 2 & 3. 226
B.7 Capture Hi-C datasets used in Chapter 2 & 3. 226
B.8 WGBS datasets used in Chapter 2 & 3. 226
B.9 RNA-seq datasets used in Chapter 2 & 3. 226
B.10 Proteomics datasets used in Chapter 2 & 3. 227
B.11 Other publicly available datasets used in Chapter 2 & 3. 227
B.12 Published ATAC-seq datasets used in Chapter 4. 228
B.13 Published mRNA-seq datasets used in Chapter 4. 228
B.14 Published microarray datasets used in Chapter 4. 228
B.15 ChIP-seq datasets used in Chapter 4. 229
B.16 Other publicly available datasets used in Chapter 4. 230

ACRONYMS

2i	Two kinase inhibitors
5-mC	5-methylcytosine
A	Anterior
AP	Anterior-posterior
ATAC-seq	Assay for Transposase-Accessible Chromatin with high throughput sequencing
auPRC	Area under precision-recall curve
auROC	Area under ROC curve
BRE	TFIIB-recognition element
BrU	Bromouridine
CAGE-seq	Cap analysis gene expression with high throughput sequencing
CHiCAGO	Capture Hi-C Analysis of Genomic Organization
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP followed by sequencing
ChIP-SICAP	ChIP with selective isolation of chromatin-associated proteins
CNN	Convolutional neural network
CPM	Counts per million
DCE	Downstream core element

DHS DNase I hypersensitive sites
DNase-seq DNase I hypersensitive sites sequencing

E Embryonic day
EpiSC Epiblast Stem Cell
eRNA Enhancer RNA
ESC Embryonic stem cells

FACS Fluorescence-activated cell sorting
FC Fold-change
FDR False discovery rate
FGF Fibroblast growth factor
FPM Fragments per million

GBM Gradient boosting machine
GRO-seq Global run on sequencing
GSK-3 Glycogen synthase kinase 3
GTF General transcription factors

H Hindbrain
H3K27ac Histone H3 lysine 27 acetylation
H3K27me3 Histone H3 lysine 27 trimethylation
H3K4me1 Histone H3 lysine 4 monomethylation
Hrep Hindbrain repressed

IDR Irreproducible discovery rate
Inr Initiator
IQR Interquartile range

KD Knock-down
KO Knock-out

LIF Leukemia inhibitory factor
LMR Lowly methylated region
LT Long-term

MAPQ Mapping quality
MCC Matthews correlation coefficient
mESC Mouse embryonic stem cells
MN Motor neuron
MPRA Massively parallel report assay
mRNA Messenger RNA

ncRNA Non-coding RNA
NES Normalised enrichment score
NFR Nucleosome free regions
NMP Neuro-mesodermal progenitor
NPC Neural progenitor cell

OE Other end
ORI Origin of replication

P-TEFb Positive transcription elongation factor
PCR Polymerase chain reaction
PIC Pre-initiation complex
POLII RNA polymerase II
PPI Protein-protein interactions
PRC2 Polycomb repressive complex 2
PTM Post translational modification
PWM Position-weight matrices

RA Retinoic acid
RAR Retinoic acid receptor

SC	Spinal cord
SCind	Spinal cord induced
scRNA	Single cell RNA sequencing
SES	Signal extraction scaling
SHH	Sonic hedge hog
SOM	Self-organizing map
STARR-seq	Self-transcribing active regulatory region sequencing
SVM	Support vector machine
TAD	Topological associated domain
TBP	TATA-binding protein
TET	Ten-eleven translocation
TF	Transcription factor
TFBS	Transcription factor binding site
TPM	Transcripts per million
TSS	Transcription start site
UMI	Unique molecular identifier
WGBS	Whole-genome bisulfite sequencing
WT	Wild type

1

Introduction

1.1 Mechanisms of transcriptional regulation

DNA encodes the information needed to generate all necessary cell types for an entire multicellular organism. The functional units of this information are encoded in the form of genes. In 1957, Francis Crick postulated the "Central Dogma of Molecular Biology" summarising the information flow in which DNA is transcribed into RNA and then later translated in proteins [1]. Resulting proteins then full fill their functions in metabolism, signalling and other cellular processes. However, although every cell in a multicellular organism contains the same genetic information, they exhibit diverse functions: a key process is how the encoded information is used at the right times and places. In 1961, Jacob and Monod published their first model of transcriptional regulation in describing the *lac* operon in *Escherichia coli* [2]. This started investigations into the field of transcriptional regulation and since then our understanding of how gene expression is regulated has continued to improve for prokaryotes as well as eukaryotes. In the first section of Chapter 1, I will present the major principles for gene transcription and the various layers of transcriptional regulation.

1.1.1 Transcription of genes to messenger RNA (mRNA)

Transcription initiation in eukaryotes

Genes are translated into messenger RNA (mRNA) in a process widely known as transcription. In a simplified eukaryotic mechanism, sequence specific transcription factors bind to promoter regions followed by chromatin remodeler recruitment and nucleosome displacement to create accessible chromatin. General transcription factors (GTF) and RNA polymerase II (POLII) then assemble at the transcription start site (TSS) of a gene to form the pre-initiation complex

(PIC) [3]. The two DNA strands are separated by disrupting the complementary base pairs to create a transcription "bubble". POLII starts to synthesise an RNA transcript of the gene until the newly synthesised molecule reaches a length of 20bp to 60bp [4, 5]. These processes together are summarised as transcription initiation.

POLII pausing and release

After transcription initiation and synthesis of a short transcript, pausing factors including DRB sensitivity inducing factor (DSIF) and negative elongation factor (NELF) bind to POLII and the carboxy-terminal heptapeptide repeat domain (CTD) is phosphorylated [6]. All genes independent of expression levels and even poised genes such as heat shock factors exhibit POLII promoter proximal pausing [7]. POLII pausing has been associated with diverse functions including: 1) formation of chromatin primed for transcription, 2) rapid induction of gene transcription, 3) as checkpoint in transcript elongation and 4) response to signalling pathways [8].
POLII can be released by direct or indirect transcription factor (TF)-mediated recruitment of a positive transcription elongation factor (P-TEFb) kinase, which in turn phosphorylates the pausing factors as well as the POLII CTD [9]. The phosphorylated NELF dissociates and the phosphorylated DSIF becomes an elongation promoting factor [10]. These events lead to POLII pause release.

Transcription elongation and termination

The phosphorylated POLII and DSIF complex continues productive elongation along the template strand to transcribe the gene into RNA. Productive RNA synthesis continues until a terminator region is reached and termination is induced by co-transcriptional RNA processing [11]. The 3'-end cleavage and polyadenylation (CPA) complex is recruited by the polyadenylation signal of the nascent transcript, leading to necessary mRNA modifications and ultimately to mRNA release and transcriptional termination [12].

1.1.2 Regulatory elements encode information to regulate gene expression

Spatiotemporal regulation of gene expression is necessary given that cells contain the same genetic information but exhibit widely different morphologies and functions. Gene transcription is regulated by the interplay of the two major classes of regulatory elements: promoters and transcriptional enhancers.

Promoter elements

Promoter elements are regions proximal to the transcriptional start site (TSS) of coding and non-coding genes regulating the recruitment and assembly of the transcriptional machinery [3, 13]. These elements can be subdivided into two regions, the so-called core-promoter and the proximal promoter [14]. The core-promoter is defined as region $\pm$50bp around the TSS containing the necessary information for POLII recruitment and transcriptional initiation. The core-promoter sequence alone results in very low basal gene expression levels [3]. In contrast, the proximal promoter is enriched for TF motifs and therefore contributes to transcriptional activation, in a similar way to enhancer elements [14].

Transcription at promoter elements Transcription initiation takes place at core-promoter by assembling of the transcriptional machinery, as described above (see section Transcription initiation in eukaryotes). Usually the recruited transcriptional machinery starts transcription in opposing directions resulting in a stable mRNA transcribed from the sense strand and an unstable none-coding RNA (ncRNA) from the antisense strand, leading to an imbalanced bidirectional expression signal [15, 16]. However, some previous literature has described promoter elements displaying productive bidirectional expression resulting in stable RNAs from both strands [17].

Core-promoter sequence motifs Base-pair resolution mapping of transcriptional initiation events using techniques such as Cap Analysis of Gene Expression (CAGE-seq) revealed enriched core-promoter sequence motifs and other chromatin features [18, 19, 13]. These motifs were associated with binding of GTFs and therefore considered to facilitate PIC assembly. Moreover, these motifs exhibit always a certain spacing relative to the TSS (reviewed in [13, 20]). The best known core-promoter motifs relevant in mammalian genomes include:

- TATA box: -31bp to -24bp relative to the TSS; bound by the TATA-box-binding protein (TBP) [21, 22, 19].
- Initiator (Inr): -3bp to +3bp around the TSS [19, 23].
- TFIIB recognition elements upstream (BREu): -38bp to -32bp relative to the TSS; bound by TFIIB [24].
- TFIIB recognition elements downstream (BREd): -23bp to -17bp relative to the TSS; bound by TFIIB [25].

- Downstream core element 1 (DCE1): +6bp to +11bp relative to the TSS; bound by TAF1 [26].
- Downstream core element 2 (DCE2): +16bp to +21bp relative to the TSS; bound by TAF1 [26].
- Downstream core element 3 (DCE2): +30bp to +34bp relative to the TSS; bound by TAF1 [26].

An example of how core-promoter motifs might impact transcription initiation is by specific recruitment of the GTF TBP to a fixed position at the promoter resulting in sharp initiation events at the TSS [27]. There are other known core-promoter motifs with less stringent spacing and as yet unclear functions (reviewed in [20]). In addition to core-promoter motifs, the local dinucleotide content of core-promoters might have an impact on transcriptional initiation as well. A lot of core-promoters overlap with regions of high CpG density called CpG islands [28]. Moreover, some promoters display a periodic pattern of A/T dinucleotides downstream of their TSS, potentially impacting nucleosome positioning at these regions [29].

Biologically relevant promoter classes Broadly mammalian core-promoters can be distinguished into three distinct classes displaying distinct sequence and chromatin features resulting in different types of initiation events (proposed by [13]). These core-promoter classes include:

- Housekeeping gene promoters: CpG island core-promoter; phased nucleosomes; active transcription marked by H3K4me3 and H3K27ac; broad transcriptional initiation pattern [30, 31].
- Developmental gene promoter: CpG island core-promoter and downstream of TSS; phased nucleosomes; active transcription marked by H3K4me3 and H3K27ac; can be Polycomb repressed (H3K27me3); broad transcriptional initiation pattern [29].
- Tissue specific gene promoter: TATA-box and initiator motifs; unphased nucleosomes; active transcription marked by H3K4me3 and H3K27ac marks activity; sharp transcriptional initiation pattern [30, 31].

Transcriptional enhancers

Enhancers were initially discovered in 1981, when Banerji *et al* described the intrinsic sequence property of the SV40 viral enhancers to enhance transcription in an episomal

reporter assay [32]. Enhancers are relatively short self-transcribing sequences (100bp-1kb), which can regulate expression of their target genes independently of location, distance and direction [33]. The binding of TFs to their sequence motifs within enhancers and the subsequent recruitment of co-activators/repressors tightly regulate the activity and function of enhancers (reviewed in [34]).

Enhancer transcription Genome-wide measurements of transcriptional initiation events revealed that active enhancers display bidirectional expression of short enhancer RNAs (eRNAs) [35, 36]. Enhancer transcription takes place at nucleosome free regions (NFR) and resulting eRNAs are short unstable RNAs which are rapidly degraded by the exosome. Interestingly, recent publications showed that eRNA expression is correlated with target gene expression [37]. However, it is unclear whether the majority of eRNAs have any biological function and if how these functions might be mediated (reviewed in [38]). Recent studies reported mechanisms how eRNAs could impact transcriptional regulation including: TF trapping, maintenance of chromatin accessibility or being involved in phase separation [39, 40, 41].

Models of enhancer architecture and grammar Attempts have been made to decipher and describe the functional organisation of TF binding sides within enhancer elements mediating enhancer activity. This "enhancer grammar" includes factors such as the number, spacing, orientation, order and affinity of TF binding sites. Different models of enhancer architecture have been suggested to describe the mechanism of TF action, including the enhanceosome, billboard and TF collective models (reviewed in [42]). The most stringent is the enhanceosome model, which assumes a tightly constrained binding site grammar and enhancer activity only if all necessary TFs are bound in the correct proportions, locations and orientation [43]. On the other hand, the billboard model proposes a flexible enhancer syntax in which TF binding mediates enhancer activity in an additive manner regardless of location [44]. Both of these models only consider sequence specific TF binding. This assumption is dropped in the TF collective model, an extension of the billboard model, in which TFs can also be recruited via protein-protein interactions (PPI) to modulate activity [45]. A recently published review summarised these models and existing evidence by suggesting a "mixed modes" model within each enhancer, where the previously described aspects of the different models could appear [42].

Enhancer core-promoter specificity A recently described property of enhancers and core-promoters is that not all transcriptional enhancers can facilitate gene expression enhancement from all types of core-promoters [46]. Measuring enhancer activity in an episomal reporter assay using different core-promoters from housekeeping and developmental genes revealed differential responsiveness of transcriptional enhancers [47]. Moreover, screening short fragments for their ability to initiate transcription in a similar setup but fixing the enhancer to either a housekeeping or developmental one confirmed these observations [48]. This suggested that enhancer core-promoter specificity seems to be encoded in the DNA sequence of these elements. This specificity is mediated by differential TF binding and subsequent recruitment of co-factors necessary for transcriptional activation. A recent publication showed that core-promoters for housekeeping and developmental genes rely on different co-factors for their activation [49].

Mechanisms of transcriptional enhancement Gene expression happens in waves of transcriptional bursts, meaning that RNA is produced in stochastic cycles of activity and inactivity [50]. Increasing transcriptional output from these bursts can be achieved by increasing the amplitude or the frequency of bursts. Previous studies in yeast showed that the information for the burst amplitude is encoded in the DNA sequence of the core promoter [51]. These findings were extended to mammalian promoter sequences by two recently published studies: the first one using single-molecule RNA microscopy to investigate HIV-1 RNA transcription in human cells and the second one confirming these observations transcriptome-wide using allele-specific scRNA-seq from human and mouse cells [52, 53]. In contrast, multiple recent studies showed that transcriptional enhancers encode the information to increase the bursting frequency. Amongst these studies, live imaging of Drosophila embryos, single-molecule RNA FISH combined with forced enhancer looping in the β-globin loci during mouse erythroid maturation and allele specific scRNA-seq from mammalian cells confirmed this mode of action [54, 55, 53].

Enhancer subtype: Super-enhancer/stretch enhancers Super-enhancers or stretched enhancers are defined as enhancer clusters making up large domains, usually spanning several kilobasepairs [56, 57]. Genome-wide super-enhancer calling is performed by aggregating either peaks from TF, Med1 or H3K27ac ChIP-seq within a predefined distance (typically 12.5kb) into domains. These larger domains are ranked based on their aggregated ChIP-seq signal, resulting in an elbow plot. Super-enhancers are then selected as those domains whose ChIP-seq signal is more extreme than the inflection point. In subsequent studies, broad

enrichment of H3K27ac was generally accepted as the major hallmark for their definition. Previous studies showed that super-enhancers are associated with genes which are crucial for cell identity or disease related, especially TFs. This concept was extended to core-regulatory circuits (CRC), in which master-regulators of cell-identity regulate each other via super-enhancers [58]. However, the concept of super-enhancers remains controversial in the field [59]. Recent studies could show that deleting sub-parts of these super-enhancers will have similar effects on target gene expression as deleting "normal" enhancer elements [60, 61].

1.1.3 Transcription factors and co-factors

Regulatory regions such as promoters and enhancers are enriched for short recognition sequences (motifs) which are bound by TFs [62]. As mentioned above, DNA-binding TFs play an essential role in regulating gene expression by helping to integrate upstream information from signalling pathways during development, tissue homoeostasis and normal cell function. A recent review reported that more than 1,600 TFs are encoded in the human genome, thus providing the combinatorial possibilities for a high degree of complexity in gene expression regulation [63]. A common feature of all TFs is that these proteins contain at least one DNA binding domain (DBD) enabling them to recognise their DNA sequence binding motif [64]. There are roughly $\sim$100 known DBDs in eukaryotes which can be used to classify TFs into families: among the largest TF families are C2H2 zinc fingers (number of genes in human, n=747), homeodomains (n=196) and basic helix-loop-helix (bHLH; n=108) [63].

Most TFs also contain additionally at least one so called effector domain responsible for their actual impact on gene regulation. These effector domains contribute in various ways to TF function including: 1) regulating TF activity (e.g. via ligand binding), 2) has enzymatic activity (e.g. KRAB or BTB domain) or 3) establishing PPIs (e.g. recruitment of co-factors) [65, 66, 67]. The effector domain is not necessarily a separate entity from the DBD, for instance in the case of the bHLH family.

Some TFs like GTFs are to directly recruit POLII and associated factors to assemble the PIC and initiate transcription [68]. However, most TFs recruit co-factors with activating or repressing functions [34]. These co-factors are often epigenetic writer enzymes (e.g. P300), chromatin remodeling complexes (e.g. SWI/SNF complex), enzymes that modify TFs or POLII (e.g. CDK9), or proteins mediating TF-POLII interactions (such as mediator proteins) [69, 70, 34]. Some of the recruited cofactors might work cooperatively to activate or increase transcription e.g. CDKs and P300 can both modify the PIC by either phosphorylation or acetylation [8]. However, whether TFs and co-factors have an activating or repressing effect

on transcription depends on the enhancer sequence context these factors bind [71]. Moreover, it was shown that co-factors have an intrinsic core-promoter compatibility and activity of these factors strongly depend on the core-promoter sequence context [49]. It is still unclear how many TF co-factors actually exist but Lambert *et al* mention that at least 443 chromatin modifying proteins were previously catalogued [63]. Further, it still remains unknown how exactly co-factors and their modification contribute to activation and repression of gene expression.

Another effect some TFs can mediate is nucleosome displacement leading to accessible chromatin for transcriptional activity at promoters and distal regulatory regions [72]. These so called pioneer factors include examples such as FOXA1 and the stem cell reprogramming factors OCT4, SOX2, KLF4 and MYC [73, 74]. A common feature of pioneer factors is their ability to bind to their sequence motif in nucleosome wrapped DNA. Pioneer factors bound to nucleosomes can mediate their displacement by actively excluding linker histones (e.g. FOXA1) or recruitment of co-factors such as chromatin remodelers (e.g. OCT4) [75, 76].

1.1.4 Chromatin structure and Histone PTMs

DNA in the nucleus of a cell is packaged into protein-DNA complexes called nucleosomes. Roughly 147bp of DNA are wrapped in one nucleosome and neighbouring nucleosomes are interspersed with linker sequences, which results in genome-wide nucleosomes spacing of 200bp$\pm$40bp [77, 78, 79]. This makes DNA more compact but also enables various possibilities for gene regulation. For example, nucleosomes physically block engagement of the transcription machinery and therefore inhibit gene expression. Each of these nucleosomes is made up of eight histone subunits including always two proteins of H2A, H2B, H3 and H4 histones [80]. These histone subunits have side changes also called tails which can be modified by epigenetic writer enzymes (TF co-factors) [81]. Among the known histone post-translational modifications (PTMs) are ADP ribosylation, acetylation, butyrylation, deamination, methylation, phosphorylation, propionylation, ubiquitinylation and sumoylation (reviewed in [79]). These modifications can directly impact chromatin structure and each modification can potentially regulate either DNA-nucleosome or nucleosome-nucleosome interactions. This in turn results in either more compact and inaccessible chromatin (heterochromatin) or in looser and more accessible chromatin (euchromatin) [82]. Beyond their role in regulating DNA accessibility or nucleosome properties, various histone PTMs have been associated with particular genomic regions (e.g. promoters or enhancers) and their function. A few of the best known histone PTMs are shown in Table 1.1. These histone

PTMs can be used as proxies to define regulatory regions such as enhancers and will be further discussed in the technical sections of this chapter.

Table 1.1: List of relevant histone PTMs and their associated functions.

Histone PTM	Genome annotation	Function	Epigenetic writer	Ref
H3K4me1	Enhancer	Active euchromatin	SET1A/B, MLL1-4	[83]
H3K4me3	Promoter	Active euchromatin	SET1A/B, MLL1-2	[84]
H3K9ac	Enhancer	Transcriptional activation	ELP3, KAT2A	[85]
H3K9me3	Broader domains (Promoter, Enhancer, Genes)	Heterochromatin	SUV39H1/2 (KMT1A/B)	[86]
H3K27ac	Promoter, Enhancer	Transcriptional activation	P300	[83]
H3K27me3	Broader domains (Promoter, Enhancer, Genes)	Polycomb repression	EZH2	[87]
H3K36me3	Gene body	Transcriptional elongation	SETD2	[87]

1.1.5 Role of DNA modifications in transcriptional regulation

Another layer of transcriptional control is encoded by covalent modifications of DNA nucleotides; for example, 5' methylation of cytosine (5-mC). To this day, 5-mC remains the best studied DNA modification, especially in the context of a cytosine followed by a guanine (CpG) [88, 89]. Most CpG dinucleotides are methylated in mammalian genomes and are locally hypomethylated at active regulatory elements. Although CpG density varies throughout genomes but they are particularly enriched in so called CpG islands, which tend to overlap promoter regions of housekeeping and developmental genes [90, 13]. These CpG islands are usually hypomethylated if the gene is actively transcribed. CpG density at distal regulatory regions is low but where they exist, they are also hypomethylated when regulatory active [91]. However, not all cells show genome-wide methylation; for example primordial germ cells, naive embryonic stem cells and cells in the pre-implantation embryo are exemplified by genome-wide hypomethylation.

CpG methylation is established by epigenetic writer enzymes of the DNA methyltransferase (DNMT) family. This family include three different DNMTs with different functions:

DNMT1 (5-mC maintenance), DNMT3A (de novo 5-mC) and DNMT3B (de novo 5-mC). 5-mC is actively erased by ten eleven translocation (TET) enzymes which oxidise 5-mC in sequential steps, initially to 5'-hydroxymethylcytosine, then to 5'-formylcytosine and finally to 5'carboxylcytosine [92]. These three intermediate DNA modifications might have distinct functions in transcriptional regulation.
Many roles have been suggested for CpG methylation in transcriptional regulation including the silencing of promoters and repeats, regulation of TF binding or binding of CTCF to insulators, recruitment of other epigenetic enzymes as well as Methyl-CpG-binding domain (MBD) proteins [93, 94, 95].
Besides of 5-mC and its associated oxidised intermediates, there are other DNA modifications such as N^6-carbamoylmethyladenine with still unknown roles in transcriptional regulation (reviewed in [96, 97]).

1.1.6 3D genome architecture and promoter-enhancer interactions

Chromatin exhibits a complex 3D organisation in the nucleus which has been associated with important functions in gene regulation; e.g. establishing chromatin interactions between enhancers and their target promoters. It was shown that the 3D chromatin architecture is cell type specific, thus providing additional evidence for its important role in gene regulation [98].
The 3D genome architecture consists of a hierarchy of organisation, including from largest to smallest: compartments, topological associated domains (TADs), sub-TADs, insulated neighbourhoods/chromatin-loops and enhancer-promoter interactions (reviewed in [99]). These structures are established by sequence specific DNA binding proteins such as CTCF or YY1 and structural proteins such as Cohesin [100, 101]. Compartments are distinguished into A and B where the A compartment is broadly transcriptional active and the B compartment largely transcriptional inactive [102]. TADs are substructures within compartments which are formed between two insulator elements bound by CTCF [103]. Smaller subTADs and insulated neighbourhoods are located within TADs and most promoter-enhancer interactions take place at this scale [104]. TAD boundaries appear to restrict the search space of enhancer elements for interactions and thereby facilitate frequent promoter-enhancer interactions. However, the gene promoter in proximity of an enhancer might not necessarily be the correct target gene and more complex modes of interactions are possible within a TAD [105]. Structures within TADs which further restrict potential promoter-enhancer interactions are called insulated neighbourhoods [104, 106]. These neighbourhoods are formed by two CTCF bound sites together with Cohesin isolating a stretch of DNA in a chromatin loop. Insulated

neighbourhoods show enrichment of super enhancers and important developmental regulators as well as tissue specific genes.

Enhancer mediated regulation of gene transcription is most likely mediated via physical contact of enhancer with their target promoter. Recent studies showed that enhancer-promoter interactions induce gene expression and therefore not a consequence of gene expression itself [107, 108]. One possibility how these promoter-enhancer interactions are formed is by binding or recruitment of YY1 at promoters and enhancers followed by Cohesin mediated loop formation [101]. However, there are examples of enhancers regulating expression from their target promoter without physically interacting such as the *Shh* gene [109].

1.2 Sequencing methods in regulatory genomics and their computational analysis

The advent of high-throughput sequencing enabled the development of various methods to measure the different layers of transcriptional regulation such as chromatin accessibility and TF binding on a genome-wide scale. In this section, I introduce and discuss the relevant functional genomic profiling methods and their associated computational analyses but in different depth and length depending on their significance for this thesis.

1.2.1 Profiling of TF binding and Histone modifications

A central question in regulatory genomics is to identify genomic regions bound by a protein of interest, such as TFs, epigenetic writers or histones carrying a specific post-translational modification (PTM) [110]. This class of methods usually relies either on antibodies that are specific to the protein of interest or alternatively tagging the protein with an epitope (e.g. FLAG-tag) [110].

Chromatin immunoprecipitation sequencing

The most widely used method based on this principle is Chromatin Immunoprecipitation combined with sequencing (ChIP-seq) [111, 110]. Initially, proteins and DNA are cross-linked using formaldehyde to fix proteins at their endogenous target loci, followed by fragmentation using sonication. The resulting short, protein-bound DNA fragments (200-600bp in length) are incubated with an antibody specific to the target of interest. Protein-bound fragments are pulled-down using immunoprecipitation (IP) followed by reversal of the cross-link. Purified DNA fragments are then used to prepare a sequencing library, usually including a size selection step for fragments ($<$300bp), which is then sequenced.
Various different methods have been developed to overcome ChIP-seq artefacts and limitations. One artefact is introduced by the formaldehyde treatment since it might also cross-link proteins in the vicinity but not directly binding DNA, leading to false-positive binding sites [112]. To overcome this problem, variations of the standard protocol have been developed using alternative cross-linking reagents or no cross-linking at all [113, 114].
Another problem is the requirement of large amounts of input DNA, making it hard to apply to primary or rare cell populations. Various methods have been developed based on different ideas such as sample pooling and indexing (iChIP) or combining with Tn5 transposase-based library preparation (ChIPmentation) [115, 116].

Another problem is the relatively low resolution, meaning that binding site peaks are generally hundreds of base-pairs long, making it difficult to identify the exact location of binding. To overcome this issue, ChIP-seq was combined with exonuclease digestion of unbound DNA to achieve motif level resolution, resulting in various new methods such as ChIP-exo or ChIP-nexus [117, 118].

Cleavage under targets and release using nuclease (CUT&RUN)

To overcome many of the above discussed limitations, a recent publication described a method called "Cleavage under targets and release using nuclease" (CUT&RUN)[119]. Briefly, nuclei are extracted, immobilised on magnetic beads and then incubated with antibody and protein A-MNase (pA-MN). Both proteins diffuse into the nuclei, antibodies bind the target of interest, pA-MN binds to the antibody and upon Ca^{++} treatment digests the bound DNA to either side of the target protein. The resulting short DNA fragments bound by the targeted protein diffuse out of the nuclei. The DNA-protein fragments are isolated into the supernatant by centrifugation, from where it can be purified and used for sequencing library preparation. The method overcomes many of the problems encountered with the standard ChIP-seq protocol including formaldehyde cross-linking artefacts and limitations related to the amount of input material. In fact, CUT&RUN was recently used to profile TF binding in single cells [120].

Computational analysis of ChIP-seq and CUT&RUN

An analysis step that most high-throughput sequencing experiments have in common is the mapping of short sequencing reads to a reference genome. Many tools have been developed based on different algorithms such as the burrows-wheeler transformation or hashing (reviewed in [121]). For ChIP-seq the most popular alignment tool remains Bowtie2 [122].

ChIP-seq experiments can yield two signal types depending on the targeted proteins: 1) sharp peaks representing TF binding and narrow histone modifications such as H3K4me1, H3K4me3, H3K27ac or 2) broad signals from ChIP against histone PTMs like H3K27me3 or H3K9me3. These different signal types need to be considered in the analysis of ChIP-seq data.

The mapped ChIP-seq reads can be used to generate genome-wide normalised signal tracks. For ChIP-seq, the choice for a normalisation method depends on the experimental setup. If there is no control or input sample available, the ChIP-seq signal is library size normalised to

counts per million (CPM) which is defined as $CPM_i = 10^6 \cdot \frac{n_i}{N}$ where n is the number of reads at position or within peak i and N is the library size. The literature also refers to the CPM also as reads per million (RPM) or in case of paired-end reads fragments per million (FPM). If a control or input experiment is available, there are multiple ways to normalise. For ChIP-seq experiments with broad enrichment signals the favoured approach is to normalise input and ChIP-seq samples to CPMs and then to compute the log2 fold-enrichment of the ChIP signal over the input. For ChIP-seq experiments with sharp peaks, the suggested method is called signal extraction scaling (SES) which initially normalises the input and ChIP-seq samples using a scaling factor derived from a cumulative coverage distribution and then computing the enrichment of ChIP signal over input [123]. The resulting signal tracks can be used for visualisation purposes as implemented in DeepTools [124].
Another important step during ChIP-seq analysis is the genome-wide search for "peaks" of read enriched regions representing TF binding or histone occurrence. Initially all peak callers define putative peak regions either as genome-wide bins or by identifying candidate peak regions by modelling the shift size. In the next step, candidate peaks are tested for statistical significance using either a Poisson test (e.g. MACS2) or Binomial test (e.g. MUSIC) [125, 126]. Peak calling is an essential analysis step and much literature has been published about potential obstacles and ideal peak calling strategies, for in depth discussion see [126]. Significant enriched peaks can be used for various different downstream analysis steps for example de novo motif discovery (e.g. HOMER) or gene ontology enrichment (e.g. GREAT) [127, 128].

1.2.2 Chromatin accessibility profiling

Chromatin accessibility can be used as a proxy to define the precise genomic location of regulatory elements in the genome. Thus, many methods have been developed based on different experimental procedures to profile chromatin accessibility such as digestion using DNase I called DNase-seq, formaldehyde assisted isolation of regulator regions (FAIRE-seq) and assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) [129, 130, 131].

1.2.3 Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)

ATAC-seq has rapidly become the most widely used approach in this category and data from ATAC-seq experiments in different systems is analysed throughout this thesis [131, 132].

Briefly, cells or isolated nuclei are incubated with hyperactive Tn5 transposases loaded with sequencing primers. The transposase can insert these primers into accessible open-chromatin regions but not densely packed heterochromatin. The resulting cleaved fragments are used to prepare a sequencing library, which represent open-chromatin regions. The method enables time-efficient and sensitive profiling of chromatin accessibility using low input material (500 cells) and even single cells [133].

Computational analysis of ATAC-seq

The first ATAC-seq analysis steps are similar to ChIP-seq pipelines, including read mapping, normalisation (usually as CPM) and peak calling. Open-chromatin peaks can be used for downstream analyses such as motif and GO enrichment analysis. There are also ATAC-seq specific analyses including TF footprinting and nucleosome calling. First, TF footprinting is based on the idea that TF binding will protect the DNA from Tn5 transposase insertion; therefore, the data allows systematic search for motifs or sites which show significant depletion of Tn5 insertions. There are two ways to proceed: scanning for motifs within peaks and then testing for statistically significant depletion in insertions or searching for depleted regions first and then annotating the enriched motifs. Both have been implemented in combination with various statistical models and corrections for potential sequence preferences of the transposase [134, 135].
Nucleosome calling aims to determine precise nucleosome dyads positions and nucleosome free regions. This analysis is only possible with paired-end reads since it requires the information about fragment lengths. The most popular tool for this is called NucleoATAC and based on cross-correlating so called V-plots [136]. V-plots are heatmaps that plot the fragment center position against fragment lengths resulting in a V shaped pattern for nucleosomes. This principle in combination with cross-correlation and background subtraction can be used to determine nucleosome occupancies [136].
A detailed description of possible analysis steps and associated tools are reviewed in [132].

1.2.4 Methods for DNA methylation profiling

Another important layer of transcriptional regulation is encoded by the covalent modification of nucleotide bases (reviewed in [88, 89]). The best studied is the methylation of CpG dinucleotides (5-methylcytosine, 5-mC) since many methods have been developed to determine the precise location of this modification. Methods for DNA methylation profiling can be divided into three major classes: 1) 5-mC specific antibodies (e.g. methylated DNA

immunoprecipitation, MeDIP), 2) enzymatic profiling (e.g. Nucleosome Occupancy and Methylome sequencing, NOMe-seq) and 3) bisulfite conversion (e,g, whole-genome bisulfite sequencing, WGBS) [137, 138, 91]. I used data from WGBS in this thesis.

Whole-genome bisulfite sequencing (WGBS)

Bisulfite treatment of DNA leads to the deamination of unmethylated cytosines into uracil [91]. Methylated cytosines are however unaffected by this treatment. These bisulfite conversions can then be measured using microarrays or more recently sequencing methods. The proportion of unconverted cytosines at a particular genomic position represents its methylation level. Thus, WGBS enables quantitative genome-wide profiling of DNA methylation levels at single-molecule resolution.

Computational analysis of WGBS data

In contrast to the above analysis methods, WGBS reads need to be mapped to a reference genome, considering mutations introduced by bisulfite treatment. A popular mapping tool is Bismark, which is based on Bowtie [139]. Mismatches caused by the bisulfite treatment can be quantified after the mapping and are interpreted as unmethylated cytosines. This information can then be used to compute methylation levels as the ratio of 5-mC over total read coverage at a given genomic position. The resulting base-pair resolution methylation maps can be used to define active regulatory regions. This can be achieved using MethylSeekR which implements a Hidden Markov model to segment the genome into methylated and partially methylated regions (PMRs). These PMRs can be further distinguished into unmethylated (UMRs) and lowly methylated regions (LMRs) [140].

1.2.5 Detection of Chromatin-Chromatin interactions

The genome forms complex spatial structures such as loops inside the nucleus. These structures can enable transcriptional regulation by bringing an enhancer into physical contact with its target promoter (reviewed in [141]). Interactions can be investigated using techniques based on the idea of chromosome conformation capture (3C) [142]. Briefly, the principles underlying 3C are: 1) formaldehyde cross-linking of DNA-DNA interactions, 2) digestion of the DNA using restriction enzymes (e.g. HindIII or MboI), 3) ligation of the ends of the digested fragments in close proximity using T4 ligase, 4) reversal of cross-linking and 5) library preparation from the proximity ligated DNA fragments [143]. Based on these

principles many variants to profile chromatin structure have been developed including 4C-seq, 5C-Seq, Hi-C, ChIA-PET, Hi-ChIP and capture Hi-C (reviewed in [143]).

Capture Hi-C

Capture Hi-C was developed to specifically profile interactions between distal regulatory regions and their target promoters [144]. This method, based on the 3C technology, uses oligonucleotides to capture specific genomic regions such as promoters or DNase hypersensitive sites (DHS) and therefore enriches for interactions of these regions [145, 146, 141]. The ligated fragments comprise the bait site and its unbaited interaction partner; thus it is crucial to perform paired-end sequencing to uncover the location of both interaction partners.

Computational analysis of Capture Hi-C

Paired-end reads from capture Hi-C are mapped to an *in silico* digested reference genome or a reference genome followed by filtering based on the expected fragments from *in silico* digestion, for instance implemented in Hi-C User Pipeline (HiCUP) [147]. Since the reads are paired-end, one end usually corresponds to the bait captured region and the other one to the interacting partner. The question then is to distinguish the observed interactions into true positive interactions and true negative ones which are based on artefacts such as spurious ligation. Identifying statistically significant interactions from capture Hi-C can be achieved using "Capture Hi-C Analysis of Genomic Organisation" (CHiCAGO) [148]. This method is based on a background model accounting for technical noise and "Brownian collisions" between fragments. Statistical testing is performed using a Delaporte model which integrates a negative binomial and a Poisson distribution.

1.2.6 Measuring gene expression using RNA sequencing (RNA-seq)

An important end-point measure of gene regulation is the productive output of transcription, commonly referred to as transcriptome. Many sequencing-based methods have been developed to profile mRNA levels from cell populations and but also single cells (reviewed in [149]). Most RNA sequencing protocols consist of the following major steps: 1) Isolating RNA from cells and DNA depletion, 2) selection for RNAs of interest (e.g. mRNA via poly(A) selection and ribosomal RNA depletion), 3) converting RNA to complementary DNA (cDNA) using reverse transcription and 4) library preparation using cDNA and 5) sequencing [150, 151].

Computational analysis of RNA-seq

In contrast to earlier sequencing methods, the mapping of RNA-seq data needs to account for gaps caused by splicing, which can be achieved using a tool like STAR [152]. An alternative is to perform a pseudo alignment against a reference transcriptome instead of the genome as implemented by Salmon or Kallisto [153, 154]. Expression is quantified by counting reads overlapping the gene annotation and normalised to transcripts per million (TPM) [155]. There are many other ways of normalising RNA-seq data, as implemented for example by DESeq [156]. Transcriptomic data contains much more information than gene expression, for example splicing; therefore, many other analysis steps can be performed which have been implemented in a wide variety of tools (reviewed in [149]).

1.3 Profiling nascent transcription

There is also an entire class of sequencing methods to study nascent transcription in a genome-wide manner. One advantage of these methods is that it allows the profiling of fragile RNAs such as enhancer RNA (eRNA) which is degraded by the exosome; these are comprehensively reviewed in [149]. Methods to profile nascent transcription can be classed into three categories: 1) Run-on methods based on nucleotide analogues (e.g. GRO-seq), 2) RNA Pol-II pull-down methods (e.g. NET-seq) and 3) metabolic labelling methods using nucleotide analogues (e.g. TT-seq) [7, 157, 158].

Global run-on sequencing (GRO-seq)

Global run on sequencing (GRO-seq) is a technique to profile nascent transcription based on the incorporation of 5-bromouridine 5'-triphosphate (BrU) during productive transcription [7, 159]. Briefly, nuclei are isolated from cells, endogenous nucleotides are washed out, exogenous nucleotides including BrU are added and transcription by DNA resident RNA polymerases is resumed. All RNAs produced during this transcription step will be labelled with BrU, which is used to pull-down transcripts using an antibody. These isolated nascent RNAs are reverse transcribed into cDNA and then used to prepare a library for sequencing.

Computational analysis of GRO-seq

GRO-seq reads need to be mapped to the corresponding reference genome using a short read tool such as Bowtie2 [122]. Genome-wide signal tracks are generated for each strand separately and normalised to CPM. These tracks can be generated at single base-pair resolution,

which represents the POLII position. Such base-pair resolution tracks can be used as input for discriminative regulatory element detection (dREG) to define transcriptional regulatory elements [160]. dREG peak calling is based on regression of observed GRO-seq data with a support vector machine model.

1.3.1 Comparison of different sequencing data based enhancer definitions

Despite the central role of enhancers in development and in the maintenance of normal cellular functions, determining their precise locations in larger genomes of eukaryotes is not straightforward. While many methods have been developed, none are definitive (reviewed in [161, 162, 163]).

Initial approaches were based on clusters of TF motifs, sequence conservation between species or low-throughput reporter assays (e.g. Luciferase assays) [32, 164, 165, 166]. This changed dramatically with the advent of high-throughput sequencing, resulting in a variety of different methods utilising different enhancer features (reviewed in Table 1.1). However, all of these methods have limitations and rather provide complementary lines of evidence. For instance, approaches like ChIP-seq, generate information about the frequency of assayed features within the sampled cell population and therefore implies enhancer location only. Open-chromatin profiling assays such as ATAC-seq detect all open loci in the genome including promoters, enhancers, silencers and insulators [131]. While they can potentially lead to false positives when classifying regions as putative enhancers, open-chromatin profiling has been widely used to improve our understanding of gene regulation and cell identity [167, 168, 169]. In contrast, CAGE-seq and other nascent transcription based assays provide a more quantitative measure of enhancer RNA (eRNA) abundance. Such approaches have the disadvantage that eRNAs are rapidly degraded, limiting the sensitivity of the method [36]. CAGE-seq studies of defined enhancers revealed many similarities between enhancers and promoters leading to a unified model of regulatory elements [170, 171]. More recently, studies have questioned the specificity of this approach and claimed that transcription can happen anywhere in open chromatin [172].

Fig. 1.1: Overview of sequencing methods used to define transcriptional enhancers.

Method	Description	Ref.
ChIP-seq	Profiling of histone post-translations modifications (PTMs), using H3K4me1 and absence of H3K4me3 as proxy for enhancer definition. H3K27ac as additional information to estimate enhancer activity.	[83]
ATAC-seq, DNase-seq	Detection of open chromatin regions due to accessibility for DNA cutting/inserting enzymes. Definition of enhancers by absence of an annotated protein-coding gene TSS (= distal regions).	[131]
WGBS	Determine location of low methylated regions (LMRs) which are distal to annotated gene promoters.	[140]
CAGE-seq, GRO-seq	Sequencing of the 5' end of RNAs to determine location of transcription start sites (TSS). Enhancer definition using bi-directionally expressed enhancer RNA (eRNA).	[36]
STARR-seq	High-throughput determination of enhancer activity of genomics cloned into a reporter vector downstream of a promoter, resulting in transcription of these sequences, which are then sequenced. Cell-type specific transcriptional program leads to transcription of responsive enhancers.	[173]

Nevertheless, a unique feature of *bona fide* enhancers should be the ability to drive gene expression from a promoter independent of their location and direction [33]. High-throughput episomal reporter assays such as massively parallel reporter assays (MPRA) or self-transcribing active regulatory region sequencing (STARR-seq) take advantage of this property and provide a quantitative measure of enhancer activity [174, 175, 173]. The major advantage of these assays is that measured activities depend only on the sequences exposed to the transcriptional program of interest. MPRAs focus on the screening of several thousand-candidate sequences whereas STARR-seq is able to work with entire genomes [162]. However, the original method was developed for smaller genomes such as *Drosophila* and was until recently not scalable for vertebrate genomes. A major disadvantage of the initial STARR-seq was that the bacterial origin of replication (ORI) functioned as a promoter, resulting in reduced enhancer signal [176]. This was resolved using the ORI as the core-promoter element, thus enabling genome-wide STARR-seq in vertebrates. Alongside this development, STARR-seq was combined with enrichment-based protocols, such as ChIP-seq or ATAC-seq, which addressed the issue of specificity by allowing a more targeted screen of either TF binding sites or open-chromatin regions [177, 178, 179, 180]. Another potential caveat of such assays is that

enhancer sequences are taken out of their native context. A recent published study examined the differences between episomal and chromosomal reporter assays [181]. They concluded that there were substantial differences between both assay types, such as reproducibility and activity prediction from chromatin marks. However, in the same study, a high correlation between the activities of chromosomal and episomal enhancer reporter assays was observed (Pearson correlation = 0.85; see Figure 2 of Inoue *et al* (2017)) [181].

A comprehensive discussion about enhancer and target gene identification methods is provided in [163].

1.4 Outline and aim of the thesis

In this PhD thesis I investigated enhancer mediated transcriptional regulation in the context of mESC pluripotency and early neural development using diverse regulatory genomic datasets derived from high-throughput sequencing methods.

In **Chapter 2**, I assessed which NANOG binding sites as measured by ChIP-seq correspond to *bona fide* enhancer elements, Moreover, I asked how enhancer activity is encoded in the DNA sequences of these NANOG bound enhancers. To this end, I designed an assay based on the combination of ChIP with STARR-seq to measure enhancer activity of NANOG bound sites in naive mESCs. I integrated the ChIP STARR-seq measurements with publicly available datasets to develop predictive models that uncover key determinants of enhancer activity.

In **Chapter 3**, I asked how enhancer activities switch between different cellular conditions, I used the ChIP STARR-seq data described in Chapter 2 to compare NANOG bound enhancer activities in naïve and primed mESCs. Analysis of the resulting dataset revealed differential active enhancers: integration of publicly available data uncovered chromatin features and sequence features that explaining observed differential activities.

In **Chapter 4**, I investigated how and when cells establish neural identity during early development of the nervous system along the anterior posterior axis. Together with collaborators from James Briscoe's laboratory, I profiled for open chromatin using ATAC-seq in an in vitro neural development model to define cell identities accurately. Further ATAC-seq measurements following regulatory perturbations allowed me to pinpoint the sequence of events needed to establish regional and neural cellular identities. The observations provided evidence to revise a long standing hypothesis in developmental biology concerning the order of events and uncovered how the earliest decisions to establish anterior and posterior nervous system are made.

Finally, I conclude the thesis in **Chapter 5** with a discussion of results presented in the earlier chapters.

2

Deciphering the sequence encoded regulatory logic of NANOG binding enhancers

Spatiotemporal regulation of gene expression in multicellular organisms is mediated by cell-type specific regulatory elements called transcriptional enhancers. Their activity is regulated by the coordinated binding of transcription factors (TFs) to specific DNA sequences encoded within these elements. However, how their regulatory potential is encoded in the DNA sequence is still poorly understood. Here, we aim to decipher the regulatory logic of NANOG-bound pluripotency enhancers in mouse embryonic stem cells (ESC). We identified putative enhancers bound by NANOG using chromatin immunoprecipitation and measured their regulatory activities using STARR-seq, a high throughput episomal reporter assay. The experiments revealed NANOG-bound elements with and without enhancer activity. Comparisons with known markers for enhancer activities such as H3K27ac (a known indicator of enhancer activity) and enhancer RNA expression supported this finding. Next, we applied predictive modelling techniques like gradient boosted trees and convolutional neural networks to uncover known and novel features that are predictive of enhancer activity. In doing so, we identified binding of the regulator estrogen-related receptor beta (ESRRB) as the most predictive feature. Finally, we integrated publicly available H3K27ac measurements from *Esrrb* knockout experiments; upon comparing with H3K27ac levels in WT cells, we found a depletion of the mark on NANOG-bound elements with enhancer activity, thus validating the importance of ESRRB as regulator. Taken together, our work demonstrates the potential of a framework based on combining high-throughput reporter assays with machine learning for understanding TF-mediated enhancer activities. ■

Declaration

The work presented in this chapter was a collaborative work within the Luscombe Lab. Mahmoud-Reza Rafiee performed wet lab experiments including generation of ChIP STARR-seq and NANOG ChIP-seq data.

I (= Sebastian Steinhauser) analysed all high-throughput sequencing data including ChIP STARR-seq, ChIP-seq, CUT&RUN, ChIP-exo, GRO-seq, ATAC-seq and additional data sources presented in this chapter. All figures in this chapter were made by myself and plots for them were generated with code I implemented.

Aylin Cakiroglu provided code and functions from her python library chromWave for the CNN training.

2.1 Introduction

Given the same genetic information in most cells of a multicellular organism, the observed diversity of cell types can only be achieved by tight spatiotemporal regulation of gene expression [182]. Cell-type specific transcriptional programs are established by regulatory elements called cis-regulatory elements (CREs) or enhancers. These elements are relatively short self-transcribing sequences (100bp-1kb), which can regulate expression of their target genes independently of location, distance and direction [33]. Enhancer activity is regulated by coordinated binding of transcription factors (TFs) to their target sequences and the subsequent recruitment of co-activators or repressors [34]. Despite the central role of enhancers in development and maintenance of normal cellular functions, determining their precise locations in larger genomes of eukaryotes is not straightforward. Though many methods have been developed to this end, none are definitive [162].

In this chapter, we aim to gain a better understanding of how enhancer activity is encoded by DNA sequence. For this purpose, we focus on transcriptional enhancers involved in the regulation of pluripotency gene expression in mouse embryonic stem cells (mESC) cultured in the presence of LIF and MEK/GSKβ inhibitors (2i) which is comparable with cells in the late blastocyst. An important hallmark of these enhancers is binding of the three key pluripotency TFs: OCT4, SOX2 and NANOG [183].

We used a previously published protocol for ChIP STARR-seq to determine *bona fide* enhancers amongst all NANOG binding sites (Fig. 2.1) [178]. Briefly, NANOG-bound DNA fragments were pulled down via chromatin immunoprecipitation (ChIP), cloned into STARR-seq screening vectors and transfected into ground-state mESC. Self-transcribing enhancers were quantified by sequencing their transcripts. Comparison with sequence reads from a pre-transfection vector library identified active enhancers but also revealed inactive TFBS. Inactive TFBS lacked the ability to enhance transcription from a minimal promoter, despite high levels of NANOG engagement. We performed a comprehensive comparative analysis of these two classes of NANOG binding sites to reveal distinguishing features. Finally, predictive modelling uncovered ESRRB as a major predictor of enhancer activity and TBX3/ZFP281 as a predictor for inactive TFBS.

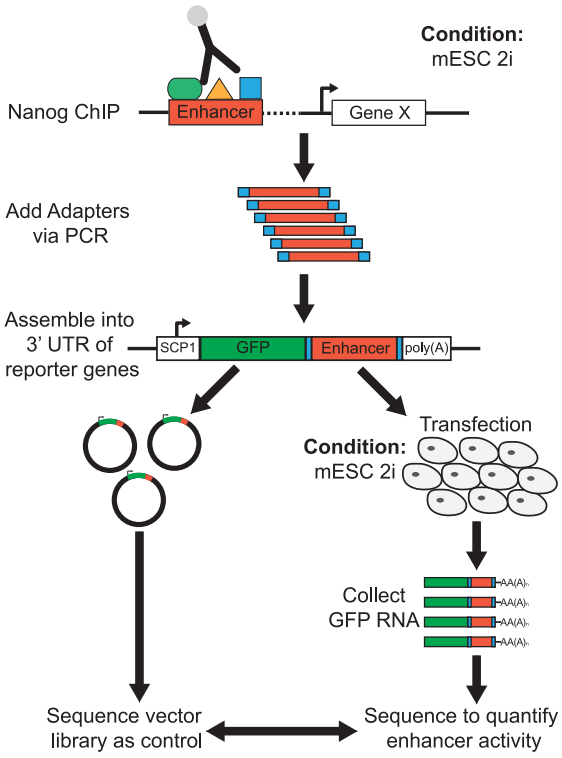


Fig. 2.1: ChIP STARR-seq experiment to determine active enhancers from NANOG binding sites. First, a chromatin immunoprecipitation (ChIP) was performed to pull down NANOG cross-linked to DNA in ground-state (2i) embryonic stem cells (ESCs). Subsequently, cross-linking was reversed, NANOG digested and associated DNA fragments were tagged with sequencing adapters via PCR. Tagged DNA fragments were cloned into the 3'UTR of GFP in the STARR-seq screening vector. In the screening vector, GFP is regulated by a SCP1 minimal promoter which is unable to induce GFP transcription without additional regulatory input. GFP expression can occur only when only if a *bona fide* enhancer element is cloned into the 3'UTR. The resulting STARR-seq vector library is transfected into ground-state mESCs and transcribed GFP mRNA is collected and prepared for sequencing. Therefore, active cloned enhancer sequences can be identified, as they will be in the 3'UTR of the expressed GFP mRNAs. To determine inactive NANOG-bound fragments, we also sequenced all the fragments cloned in the vector library. Cloned and transcribed fragments were target using the PCR added sequencing adapters [184]. We did not extract and sequence the transfected vector library. Finally, reads from control and STARR-seq libraries were aligned against *mm10* reference genome, quantified and differential analysis was performed.

2.2 Methods

2.2.1 Experimental procedures

ChIP STARR-seq libraries were prepared by Mahmoud-Reza Rafiee (Postdoc in the Luscombe Lab, The Francis Crick Institute) and sequenced by the DNA sequencing facility at the Okinawa Institute of Science and Technology. For brief description of the ChIP STARR-seq protocol see [178]. For all STARR-seq experiments performed, we sequenced fragments cloned in the vector library and the same fragments post-transfection. Cloned fragments and transcribed enhancer fragments were sequenced using the sequencing adapters added during PCR, which are specific for the i5 and i7 sequencing primers (as described in detail in [184]). We did not extract and sequence the transfected vector library. A list of newly generated ChIP-seq and STARR-seq datasets is summarised in Supplementary Table B.1 and B.4.

2.2.2 ChIP-seq pre-processing & analysis

Reads were trimmed using trim_galore (default settings) to remove poor quality bases and sequencing adapters (*https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*). Alignments were generated using bowtie2 with –sensitive as additional option and *mm10* as reference genome [122]. Only reads that mapped uniquely with high quality ($MAPQ > 30$) were retained and duplicates were removed. Read enriched regions were called using MACS2 with the following parameters –g mm –p 0.001 [185]. We provided MACS2 with an input control if one was available. Signal tracks for samples with a corresponding input control were generated using deepTools bamCompare including thee following parameters –scaleFactorsMethod SES –ratio log2 –bs 25 –ignoreDuplicates [124]. For samples without control, signal tracks were generated with deepTools bamCoverage using -bs 25 –normalizeUsing CPM –ignoreDuplicates [124]. Histone post-translational modifications as well as TF ChIP-seq signals were visualised as heatmaps using deepTools computeMatrix in combination with plotHeatmap [124]. Metaprofiles were computed using the ScoreMatrixBin function from genomation [186]. ChIP-seq samples were quantified using featureCounts functions from the Subread package with the following settings: strandSpecific=0, allowMultiOverlap=TRUE, countMultiMappingReads=FALSE, ignoreDup=TRUE [187]. Resulting counts were normalised to CPM and if input control was available log2(FC) calculated. All in-house generated and publicly available ChIP-seq datasets used in this study are listed in Supplementary Table B.4.

Definition of putative NANOG binding enhancers

ChIP-seq samples used to define putative NANOG binding enhancers are listed in Supplementary Table B.4. Peak calling was performed as described above. If replicates were available, irreproducible discovery rate (IDR) analysis was performed and results filtered with a threshold of $IDR \leq 0.05$ to create a consensus between replicates [188]. If no replicates were available, we used all called peaks on the single replicate. The resulting peaks from each condition were merged into a consensus NANOG peak set using bedtools merge and extended to $\pm 250bp$ around peak summits [189]. These peaks were further filtered for overlap with ENCODE *mm10* blacklisted regions [190]. Next, a set of non-overlapping random peaks was computed using bedtools shuffle with the following settings: -f 0 -seed 42 -noOverlapping [189]. We computed the STARR-seq vector library fragment coverage within NANOG consensus peaks and random peaks using countOverlaps from GenomicRanges [191]. A threshold to distinguish STARR-seq fragment coverage from background was computed as 99^{th} percentile of random peak coverage. This fragment count threshold was used to filter NANOG consensus peaks to obtain a high confidence, sufficiently deeply covered set of putative NANOG binding enhancers. This peak set was used in the downstream analyses presented in Chapters 2 and 3.

ChIP-seq peak enrichment analysis

Published ChIP-seq peaks for TF, chromatin associated proteins and histone modifications were obtained from LOLA and Cistrome [192, 193] (Supp. Table B.11). Enrichment analysis using these peak sets was performed using a one-sided Fisher's exact test (as implemented in the R function "fisher.test"). P-values were corrected for multiple testing utilizing the Benjamini-Hochberg procedure. The Matthews correlation coefficient (MCC) was computed from contingency tables used for testing with the Fisher's exact test (for MCC definition see [194]).

2.2.3 ChIP STARR-seq pre-processing & analysis

Fragments were trimmed to remove low quality base calls as well as sequencing adapters using trim_galore (default settings; *https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*). Trimmed fragments were aligned against the *mm10* reference-genome using bowtie2 with the following parameters -X 1000 –end-to-end –very-sensitive [122]. Subsequently unmapped, low quality mapping, multi-mapping and duplicated reads were removed. In a STARR-seq experiment, duplicated fragments can originate from enhancer

self-transcription or PCR amplification. Since we did not use unique molecular identifier (UMIs) to distinguish PCR duplicates from self-transcribed fragments, we had to rely on unique fragments [173, 178]. Signal tracks were computed as log2(FC) STARR-seq over vector library using deepTools bamCompare with the following options: –normalizeUsing CPM –operation log2 -bs 10 –scaleFactorsMethod None –ignoreDuplicates [124]. To perform the fragment based analysis, as in the STARR-seq quality control (QC) section (see Results), we imported the vector library as well as STARR-seq fragments into R as GenomicRanges using GenomicAlignments and Rsamtools [191]. Fragment abundances of the vector library and STARR-seq samples within NANOG consensus peaks (see ChIP-seq section) were quantified using the countOverlaps function from GenomicRanges. The resulting count table was used as input to edgeR to compare fragment abundance in the STARR-seq versus the vector control using the quasi-likelihood F-test [195]. Based on this analysis, we categorised NANOG consensus peaks into three classes: active enhancers ($FDR \leq 0.05 \& log2(FC) > 0$), inactive TFBS ($FDR \leq 0.05 \& log2(FC) < 0$ and grey-zone Nanog TFBS ($FDR > 0.05$). Active enhancers and inactive Nanog TFBS were annotated using the annotatePeak function from ChIPseeker with GENCODE release M14 [196, 197].

2.2.4 CUT&RUN pre-processing & analysis

Publicly available CUT&RUN datasets were downloaded as listed in Supplementary Table B.3. Low quality base-pairs and sequencing adapters were trimmed using trim_galore (default settings). Trimmed fragments were aligned to the *mm10* reference genome using bowtie2 with default parameters [122]. Aligned fragments were filtered for unmapped, multi-mapping and PCR duplicated reads. Normalised signal tracks were computed using deepTools with following settings: –normalizeUsing CPM –extendReads -bs 25 –ignoreDuplicates –maxFragmentLength 140 [124]. CUT&RUN signals within NANOG consensus peaks were quantified as counts using the featureCounts function from the Subread package with following options: strandSpecific=0, allowMultiOverlap=TRUE, countMultiMappingReads=FALSE, isPairedEnd=TRUE, ignoreDup=TRUE, maxFragLength=140 [187]. Resulting counts were library size normalised to CPM. Metaprofiles were computed using the ScoreMatrixBin function from the genomation package with default settings [186].

2.2.5 ChIP-exo pre-processing & analysis

ChIP-exo reads were trimmed, mapped and filtered as described in Section 2.2.2. Signal tracks for plus and minus strands were computed using the bamCoverage function from deep-

Tools with the following options: –normalizeUsing CPM –Offset 1 -bs 1 –ignoreDuplicates [124]. Depending on the strandedness of the signal track, the options "–samFlagInclude 16" in case of the minus strand or "–samFlagExclude 16" in case of the plus strand were used. The ChIP-exo peaks used in this chapter were downloaded alongside the raw data from the original study as listed in Supplementary Table B.2.

2.2.6 RNA-seq pre-processing & analysis

RNA-seq samples were quantified using Salmon (quasi mapping mode) with the following options: -l A –seqBias –numBootstraps 50 [154]. GENCODE release M14 was used as transcriptome annotation [197]. Quantification of single and paired-end reads were performed with the same options. The resulting counts were normalised to transcripts per million (TPM). The publicly available RNA-seq datasets used in this study are listed in Supplementary Table B.9.

2.2.7 GRO-seq pre-processing & analysis

The downloaded GRO-seq datasets are listed in Supplementary Table B.6. GRO-seq reads were aligned to the *mm10* reference genome using STAR with the following settings: –genomeLoad NoShareMemory –twopassMode Basic –outSAMStrandField intronMotif –outFilterIntronMotifs BAM SortedByCoordinate –outFilterMultimapNmax 2 –outFilterMismatchNmax 999 –outFilterMismatchNoverReadLmax 0.04 –bamRemoveDuplicatesType UniqueIdentical [152]. Strand specific signal tracks were computed using deepTools bamCoverage with –normalizeUsing CPM –ignoreDuplicates -bs 1, –samFlagExclude 16 (forward strand) or –samFlagInclude 16 (reverse strand). These signal tracks were used as input for the ScoreMatrixBin function from genomation and visualised as metaprofiles [186]. GRO-seq read abundance was quantified using featureCounts (Subread) with the options, strandSpecific=0, allowMultiOverlap=TRUE, countMultiMappingReads=FALSE, isPairedEnd=FALSE, ignoreDup=TRUE [187]. The resulting count table was library size normalised to CPM.

2.2.8 ATAC-seq pre-processing & analysis

Reads were trimmed using trim_galore (default settings) to remove sequencing adapters and low quality base calls. Trimmed reads were mapped to the *mm10* reference genome using bowtie2 with the following settings: -X 2000 –sensitive-local [122]. Aligned fragments were

filtered for low quality mapping (MAPQ < 30), unpaired fragments, chrM mapping and PCR duplicates. Peak calling was performed with MACS2 using the parameters -g mm –p 0.01 –nomodel –f BAMPE [185]. Signal tracks were generated by counts per million (CPM) normalisation using deepTools bamCoverage with the following options: -bs 1 –extendReads –ignoreDuplicates –normalizeUsing CPM [124]. For all insertion level analysis such as TF footprinting, insertion positions were adjusted by +4bp for the plus strand and -5bp for the minus strand [131]. Nucleosome positions and nucleosome free regions (NFR) were determined using NuceloATAC with default parameters and peaks called by MACS2 which were rescaled to $\pm 2500bp$ around peak summits as search space [136]. TF footprinting was performed by matching motifs to peak sequences to determine motif positions (see TF motif matching sections). Tn5 insertion for $\pm 250bp$ around motifs were computed using a custom R function and normalised to the total number of profiled insertions. The downloaded publicly available ATAC-seq datasets are listed in Supplementary Table B.5.

2.2.9 Genome browser track plots

Normalised signal tracks (bigWigs) from different sequencing experiments were visualised as genome browser track plots using pyGenomeTrack [198].

2.2.10 Capture Hi-C analysis

CHiCAGO pre-processed DHS capture Hi-C interactions for ground-state (2i) and primed (serum) ESCs were downloaded [145]. Interactions were filtered for ones with a CHiCAGO score ≥ 7 and at least 5 supporting fragments in each replicate. The bait and "other end" (OE) fragments were annotated with overlapping genes from GENCODE V14, active NANOG enhancers and inactive NANOG binding sites. Based on these annotations, we computed the following statistics: interaction type summary, CHiCAGO score as proxy for interact strength, number of interactions and interaction distance [148].
We made use of another previously published version of the same DHS capture Hi-C dataset which was well annotated with different chromatin states [199]. These interactions were used to compare active enhancers and inactive TFBS in terms of chromatin state. We performed differential enrichment analysis using a one-sided Fisher's exact test followed by multiple testing correction using the Benjamini-Hochberg procedure.
All used capture Hi-C datasets are listed in Supplementary Table B.7.

2.2.11 Peak-gene annotation and enrichment analysis

Active NANOG binding enhancers and inactive TFBS were assigned to their target genes using three different approaches. First, target genes were defined based on the annotated CHiCAGO interactions from *Joshi et al* [145]. Genes were linked to peaks for interactions within a distance of $\pm 500kb$ around peak summits and all interactions were considered independent of the peak overlapping bait or OE fragment. Second, a pre-processed, annotated publicly available high-resolution capture Hi-C interaction set, which was pre-processed and annotated was downloaded [146]. Genes were assigned to a peak if peaks were overlapping the OE fragment of an interaction. Third, peaks were annotated with nearby genes using GREAT, a tool for peak-gene annotation, with default settings (Basal plus extension) [128]. Expression of target genes were obtained as TPM values from processed RNA-seq (see RNA-seq preprocessing). Pathway enrichment was performed using EnrichR with default settings and gene sets from WikiPathway [200, 201].

2.2.12 Core-promoter element enrichment

Core-promoter element annotation for the *mm10* reference genome were downloaded from the Eukaryotic promoter database (EPD) [202] (Supp. Table B.11). CpG island annotation was downloaded from the UCSC genome browser using the table browser tool [203]. CpG islands were associated with genes by overlap with GENCODE M14 promoters ($\pm 500bp$ around GENCODE TSS) [197]. Core-promoter enrichment analysis in gene sets associated with active and inactive enhancers was performed using a one-sided Fisher's exact test followed by multiple testing correction with the Benjamini-Hochberg procedure.

2.2.13 Binary classification of active enhancers and inactive TFBS

Active enhancers and inactive TFBS were split into a training set and a hold-out test set according to their chromosomal location as follows: training set chr1-chr8, chr10, chr12-14; test set chr9, chr11; CNN validation set chr15-19, chrX, chrY. Active enhancers and inactive TFBS served as target variables to train binary classification models.

GBM feature selection, hyperparameter optimization and model training

We used a two-step feature selection procedure to identify meaningful TFs and their associated TF motifs to discriminate between active enhancers and inactive TFBS. First, we downloaded publicly available proteomics data, for NANOG co-binding proteins (ChIP-SICAP)

and ESC chromatin associated proteins (Chromatome) [204, 205]. A set of reproducible, biologically relevant NANOG co-binding TFs were chosen by overlapping detected TFs in the three different datasets (see Results 2.3.10). To compare the performance trained with our NANOG co-binding TFs, we constructed a control set comprising ESC chromatin associated TFs which did not overlap with the list of ChIP-SICAP identified NANOG co-binding TFs. In a second step, we identified TF-PWM pairs from a recently published consensus review mapping multiple PWMs to each TF [63] (Supp. Table B.11). To assign the most predictive PWM to a TF, we devised a strategy of pruning these TF-PWM links based on their predictive power for active/inactive sites. PWMs were matched to NANOG binding site sequences and resulting PWM scores were used as predictors to distinguish active enhancers from inactive TFBS. The predictive power of each PWM was evaluated based on the area under the ROC curve (auROC). TF-PWM links maximising the auROC were retained. This step was performed for both, control and NANOG co-binding TFs. Additionally, we included another control set of highly predictive features by thresholding for PWMs with $auROC > 0.6$. This feature selection procedure resulted in three distinct feature sets (NANOG co-binding TF PWMs, Control TF PWMS and predictive PWMs). We added dinucleotide frequencies to all of these three feature sets to account for potentially higher order sequence features or left out features.

Gradient boosting machine (GBM) hyperparameter search and model training was performed using the caret R-package which serves as a wrapper for the GBM package [206, 207]. For each of the feature sets we performed a grid search to optimise model hyperparameters as listed in Table 2.1. Models were trained on the training set using 10-fold cross-validation (CV) and the Matthews correlation coefficient (MCC) was tracked. Caret automatically selected the best set of hyperparameters based on the MCC to train a final model.

Table 2.1: GBM hyperparameter search space.

Hyperparameter	Variable name	Values
Maximum tree depth	interaction.depth	3,6,9
Number of trees	n.trees	100,200,300,400,500
Minimum number of observations in terminal node	n.minobsinnode	10
Learning rate	shrinkage	0.05

gkm-SVM hyperparameter optimization and model training

We used the R implementation of the gapped k-mer support vector machine (gkm-SVM) package and sequences of active enhancers as well as inactive TFBS as input [208]. Initially, we performed a grid-search to optimise hyperparameters of the gapped k-mer computation (Table 2.2). The best parameter set was chosen based on the area under precision-recall curve (auPRC) from 10-fold CV SVM training on the training set. The final SVM model was trained on the training set using these parameters.

Table 2.2: gkm-SVM hyperparameter search space.

Hyperparameter	Variable name	Values
K-mer length	L	6-14
Gap length	K	6-L
Number of informative columns	maxnmm	1-4

Convolutional neural network hyperparameter optimization and model training

Initially, NANOG peak sequences were one-hot encoded using A (1, 0, 0, 0); C (0, 1, 0, 0); G (0, 0, 1, 0); T (0, 0, 0, 1). These encoded sequences were used as input to train Convolutional neural networks (CNNs). Due to an imbalanced class distribution for our STARR-seq data, we computed class weights as: $w_y = \frac{N}{2 \cdot n_y}$ where N is the total number of sequences and n_y the number of sequences in class y. Class weights were provided for CNN training on STARR-seq data.

To determine the optimal CNN architecture, we performed hyperparameter optimization using the Tree of Parzen Estimator (TPE) algorithm as implemented in the hyperopt package [209]. The hyperparameter search space was predefined as listed in Table 2.3. Finally, the hyperparameter search was performed for 100 iterations while monitoring the MCC on the validation set as the performance measure to select the best model architecture.

This training strategy was used to train a CNN directly on our STARR-seq data. However, these CNNs were initialised with random weights which might result in the need for a larger number of training samples to reach a comparable performance as gkm-SVM or GBMs. Therefore, we made use of technique called transfer learning to decrease the amount of data needed to reach comparable performance [210]. For this purpose, we trained a CNN

to predict NANOG binding from ChIP-seq peaks using the same strategy as described above with a slightly different hyperparameter search space (Table 2.3). As input, we used sequences from NANOG ChIP-seq consensus peaks as positive examples and randomly shuffled these sequences while preserving dinucleotide frequencies to generate negative examples. Training, validation and test data was stratified as described for STARR-seq. Hyperparameter optimisation and training was performed for this ChIP-seq CNN as described above. Next, we extracted the learned weights in the first convolutional layer from this trained ChIP-seq CNN and used them to initialise a new CNN. The ChIP-seq initialised CNN was then trained on the STARR-seq as previously described.

Table 2.3: STARR-seq CNN hyperparameter search space.

Hyperparameter	STARR-seq CNN values	ChIP-seq CNN values
Number of convolutional layers	1	1
Kernel lengths	12, 16, 24	12, 16, 24
Number of convolutional filters	32, 64, 128	32, 64, 128, 256, 512
Dropout rates	0.1-0.5	0.05 0.5
L2 regularizer	0-0.01	-
Activation functions	ReLu	ReLu
Pooling	flatten, max, average	flatten, max, average
Fully connected layer	1	1
Number of neurons in fully connected layer	16, 64, 128	16, 64, 128, 256, 512, 1028
Early stopping	5, 10, 20	5, 10, 20
Reduce learning rate patience	5, 10, 20	5, 10, 20
Reduce learning rate factor	0.05-0.2	0.05-0.2
Optimiser	Adam	Adam
Learning rate	0.0001-0.005	0.000001-0.01
Batch size	64, 128, 256	128, 256, 512
Epochs	100	100
Loss	Binary cross-entropy	Binary cross-entropy

Evaluation and comparison of trained models

All trained models were evaluated on a hold-out test set using the following performance measures: area under precision recall curve (auPRC), Matthews correlation coefficient (MCC), balanced accuracy, sensitivity and specificity [194].

GBM feature importance analysis

Scaled feature importance scores for the NANOG co-binding TF GBM model were computed using the function varImp from the R-package caret [206]. varImp is a wrapper function for the relative.influence function from the gbm R-package [207]. Briefly, the relative influence of a feature is defined as loss improvement made in each split the feature was used averaged over all trees [211].

CNN feature importance analysis

For each active enhancer and inactive TFBS, base-pair resolution importance scores were computed using DeepLIFT with default parameters [212]. Resulting scores were used to perform de novo motif discovery with TF-MoDisco (default settings) [213]. Resulting de novo motifs were annotated using TomTom (additional settings: -thres 0.05) and the compareMotifs function from HOMER (default parameters) [214, 127]. Consensus annotation was obtained by manually inspecting TomTom as well as Homer outputs. Annotated TF-MoDisco motifs were clustered using STAMP with following options: -align SWU -tree UPGMA -chp -ma IR [215]. Feature importance for each de novo motif was computed as described in the second step of the feature selection procedure for GBMs resulting in a volcano like plot.

2.3 Results

In this chapter, we sought to document and analyse which of the thousands of TFBS observed in a standard ChIP-seq assay are functional transcriptional enhancer elements. To this end, we focused on the pluripotency factor NANOG and its occupied genomic binding sites. To determine if a particular NANOG binding site was a *bona fide* enhancer, we made use of a previously published method combining ChIP with the high-throughput enhancer reporter assay STARR-seq [178].

2.3.1 Defining high-confidence NANOG TFBS from ChIP-seq and ChIP STARR-seq

The first step to address our question was to define a high confidence set of putative NANOG binding enhancers. Therefore, we analysed NANOG ChIP-seq datasets from three different conditions including ground-state (2i), primed (serum) mouse embryonic stem cells (ESC) and mouse epiblast like stem cells (EpiSCs). NANOG ChIP-seq experiments from these three conditions were included to establish a broad initial search space for potential enhancers. Peak calling was performed independently for each replicate and the consensus between replicates was computed using the irreproducible discovery rate (IDR) procedure with a threshold of 0.05 [188]. NANOG ChIP-seq from 2i ESCs yielded 39,973 and 29,539 peaks for each replicate, respectively. Integration of replicates to a Nanog 2i consensus peaks set resulted in 15,651 reproducible peaks. In contrast, ChIP-seq from serum ESCs gave us for the first replicate 24,986 and the second replicate 26,299 peaks. Computing the consensus peak set for these replicates resulted in 5,168 peaks. Since only one replicate was available for the NANOG EpiSC ChIP-seq, we included all 4,715 peaks from this condition. Finally, these three different peak sets were inclusively merged by taking the union of peaks, leading to 20,321 potential NANOG bound enhancers. Importantly, this consensus NANOG peak set was also used to compare STARR-seq from 2i and serum ESCs (see Chapter 3).

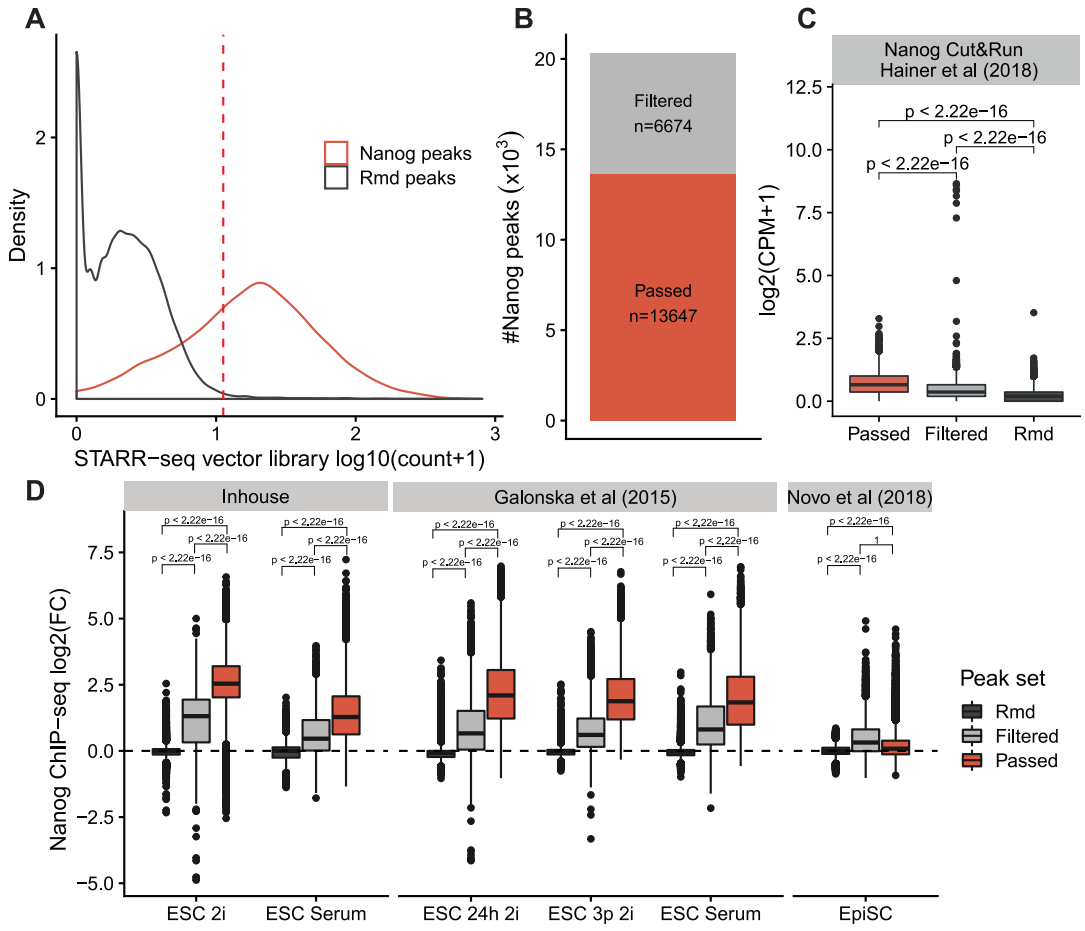


Fig. 2.2: Definition of NANOG consensus binding sites from ChIP-seq and STARR-seq vector library. (A) STARR-seq vector library coverage within NANOG binding peaks (red) and background as defined by randomly drawn genomic regions (black, "Rmd"). Dotted line indicates 99% quantile of background coverage and filtering threshold for NANOG peaks to be called as sufficiently covered by the vector library ("passed"). **(B)** Number of NANOG peaks passing the filtering step (red) and peaks filtered out due to low coverage in the STARR-seq vector library (grey). **(C)** Boxplot comparison of NANOG CUT&RUN signal within random background regions (dark grey), filtered out NANOG peaks (grey) and final set of NANOG binding peaks (red). P-values from one-sided Wilcoxon test. **(D)** Boxplot visualising NANOG ChIP-seq enrichment within random background regions (dark grey), filtered out NANOG peaks (grey) and final set of NANOG binding peaks (red). ChIP-seq datasets were obtained from three different studies as indicated in panel heading. ChIP-seq conditions include ESC 2i, ESC serum, ESC grown for 24h in 2i (ESC 24h 2i), ESC three passages 2i (ESC 3p 2i) and EpiSC. P-values from one-sided Wilcoxon test.

One important consideration is that it is only possible to detect *bona fide* transcriptional enhancers if the NANOG consensus peaks show fragment enrichment in our STARR-seq vector library. Therefore, the same number of non-overlapping random regions were sampled from the mouse genome to estimate background coverage levels (see Methods). We

defined a coverage threshold to filter our enhancer set equal to the 99^{th} percentile of the background distribution corresponding to roughly 12 unique STARR-seq vector library fragments (Fig. 2.2A). Filtering our putative Nanog binding enhancers with this threshold resulted in 67.16% (13,647/20,321) sufficiently covered NANOG peaks (Fig. 2.2B). This results in a set of 13,647 NANOG peaks represented in the STARR-seq vector library data (NANOG consensus peaks), 6,674 Nanog peaks that are not represented in the STARR-seq data (NANOG filtered out peaks) and 20,321 randomly assigned control peaks (random peaks), which we call the three peak set.

We performed several additional analysis steps to ensure that this procedure resulted in a high confidence set of potential NANOG binding enhancers. First, we compared NANOG ChIP-seq enrichment from three different studies and our dataset in five different conditions including ESC grown long-term 2i (ESC 2i), 24h 2i (ESC 24h 2i), three passage 2i (ESC 3p 2i), serum (ESC serum) and epiblast stem cells (EpiSC) [216, 217]. This analysis revealed that our set of putative NANOG binding enhancers had the highest NANOG ChIP-seq enrichment across all ESC conditions ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.2D). Though NANOG filtered out peaks also showed higher ChIP-seq enrichment than the random control regions ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test) and in EpiSC even higher median enrichment than our NANOG consensus set ($p = 1$, one-sided Wilcoxon test). This analysis provided the first evidence that our consensus peaks contained a high number of true-positive binding sites using an independent data set. ChIP-seq data contain artefacts, such as those artefacts from formaldehyde cross-linking [218]. Therefore, we performed a similar analysis using CUT&RUN data, a technique that does not rely on formaldehyde cross-linking [119]. This independently confirmed significant differences in NANOG binding signal between the three peak sets ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test). Furthermore, the CUT&RUN data exhibited the same trend as the ChIP-seq data in which our NANOG consensus peak set showed highest signal enrichment (Fig. 2.2C).

NANOG binds to specific DNA sequence motifs and their enrichment could provide additional support for true-positive binding-sites. Therefore, we checked the enrichment of NANOG motif, as well as the OCT4-SOX2-TCF-NANOG composite motif in the three peak sets. Both motifs showed significant enrichment in our NANOG consensus peak set compared with the other two peak sets ($p = 4.96 \cdot 10^{-54}$ and $p = 4.14 \cdot 10^{-173}$, one-sided Fisher's exact test). In addition, these motifs are centred on ChIP-seq peak summits and depleted in random control peaks (Fig. 2.3A). We also observed, for both motifs, the highest motif match scores for our consensus peak set; in fact, scores were significantly higher than the NANOG filtered out peaks ($p = 1.3 \cdot 10^{-4}$ and $p = 4.8 \cdot 10^{-10}$, one-sided Wilcoxon test).

Taken together, we defined a high-confidence consensus set of 13,647 putative NANOG bound enhancers which were characterised by signal in both NANOG ChIP-seq as well as CUT&RUN assays and enrichment of NANOG and OCT4-SOX2-TCF-NANOG motifs.

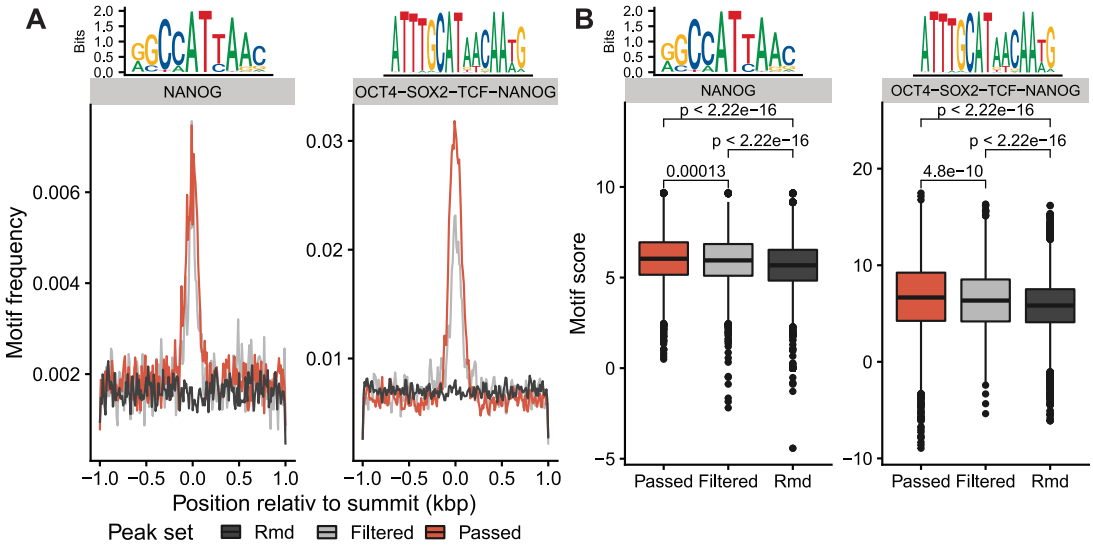


Fig. 2.3: NANOG motif enrichment for putative NANOG enhancer set. **(A)** Metaprofile summarising motif frequencies in 10bp tiles for $\pm 1kb$ around peak summits. **(B)** Motif scores for each peak visualised as boxplot. P-values from one-sided Wilcoxon test. Colour indicates peaks passing filtering (red), filtered out peaks (grey) and random background regions (dark grey). *Left panel:* NANOG consensus motif from HOMER motif library [127]. *Right panel:* OCT4-SOX2-TCF-NANOG composition motif from HOMER motif library [127]. Respective sequence logos shown on top.

2.3.2 STARR-seq quality control analysis

Having defined a set of putative NANOG binding enhancers, we performed comprehensive quality control analysis of the STARR-seq samples generated in this study. Briefly, we used our STARR-seq vector library to transfect ESC cells in two different conditions including ground-state (2i) as well as primed (serum). In Chapter 2, we will focus on the 2i dataset and in Chapter 3, we will compare 2i and serum conditions. However, the quality control of all samples was performed together and we will discuss all datasets in this section.

The STARR-seq vector library was sequenced to a read depth between 7.70 and 11.9 million fragments giving between 5.80 and 7.94 million unique fragments after duplicate removal (Fig. 2.4A). Across all conditions, our STARR-seq vector library comprised 20.50 million uniquely cloned fragments. RNA was collected post vector transfection and sequenced to a depth of between 18.61 and 19.41 million fragments resulting in 2.88 to 3.98 million

fragments after duplicate removal. The high duplication rates (between 79.5% and 84.5%), is partly explained by self-transcription of *bona fide* enhancer fragments but also PCR amplification. We chose to remove duplicates as a precautionary measure as we did not employ unique molecular identifiers (UMIs) during library preparation [219]. Examining the number of unique fragments as a function of sequencing depth confirmed that our samples were sequenced to sufficient depth to recover self-transcribing fragments (Fig. 2.4C). Additionally, we could observe a much greater fragment complexity in our sequenced vector library (more unique fragments) compared with post-transfection samples confirming purifying selection of enhancer fragments by self-transcription. A potential confounding factor is that some vectors were never transfected. Next, we compared the vector library and STARR-seq samples on the fragment level with each other. We should recover the same fragments before and after transfection if these fragments were able to enhance transcription from a minimal promoter. The fragment length distributions in the vector library and STARR-seq are similar: the vector library showed a mean fragment length of 283bp, STARR-seq 2i condition of 274bp and STARR-seq serum condition of 263bp (Fig. 2.4B). All three biological replicates from both conditions display 60.00-65.00% fragment level overlap with the vector library, which is similar genome-wide as well as in our NANOG binding enhancer set (Fig. 2.4D). The missing 35.00-40.00% could be explained by the fact that our STARR-seq vector library exhibited high complexity (20.5 million unique fragments), meaning that it would be necessary to sequence extremely deeply to recover all self-transcribing fragments (Fig. 2.4C). In contrast, 53.7% of vector library fragments were not discovered in any of our STARR-seq samples: it is possible that the majority of cloned fragments might not be able to enhancer transcription above basal levels or was not transfected. Next, we compared as a measure for fragment enrichment within NANOG peaks the so called "fraction of reads in peaks" (FRIP) between conditions as well as replicates (Fig. 2.4A left panel). We observed similar ranges of FRIPs for our STARR-seq samples: 13.40%-13.90% for our ESC 2i and 11.80%-12.30% for ESC serum condition. By contrast, our STARR-seq vector library samples showed FRIPs of 8.40% which again confirmed enrichment of self-transcribing fragments within NANOG peak regions via the transfection step (Fig. 2.4A left panel).

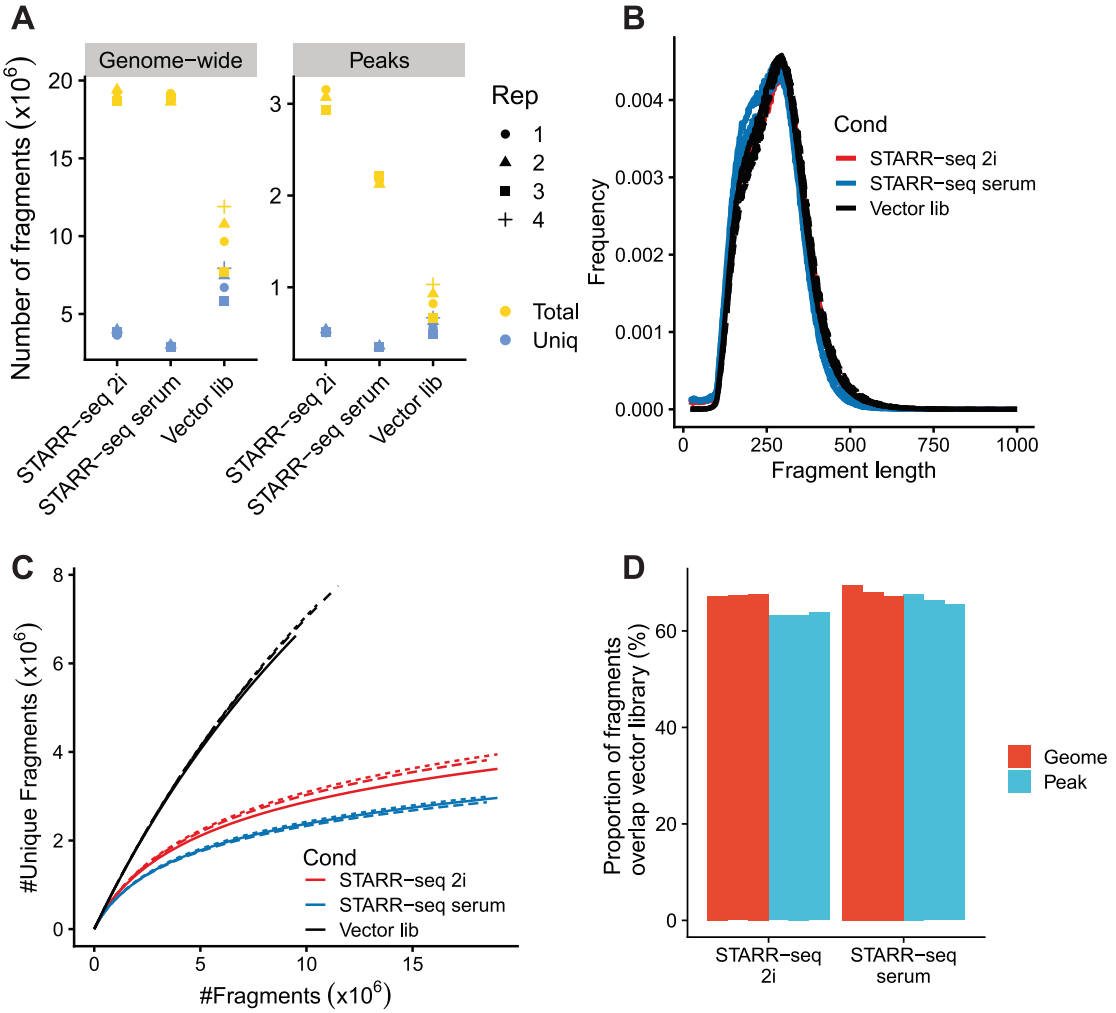


Fig. 2.4: Quality control analysis of vector library and STARR-seq samples (A) Number of unique (blue) and duplicated (yellow) fragments sequenced for vector library and STARR-seq samples. *Left panel:* Genome-wide comparison of the fragment numbers. *Right panel:* Fragment numbers within NANOG consensus peaks. Shape indicates replicate number. **(B)** Fragment length distributions for vector library (black) and STARR-seq 2i (red) and serum condition (darkblue). Linetype highlights different replicates. **(C)** Number of unique fragments as a function of sequence depth. Comparison between vector library (black) and STARR-seq 2i (red) and serum (darkblue) samples. Linetype highlights different replicates. **(D)** Number of precisely recovered vector library fragments in STARR-seq. Colour indicates comparison level either genome-wide (red) or within NANOG consensus peaks (blue).

As additional quality control, we compared ChIP-seq samples, STARR-seq vector library (NANOG ChIP cloned into vectors) and STARR-seq samples. Comparison between NANOG ChIP-seq and the vector library revealed similar NANOG binding strengths within our NANOG consensus peak set ($R = 0.62$, $p < 2.2 \cdot 10^{-16}$, Correlation test; Fig. 2.5A). Com-

parison of the STARR-seq vector library with one replicate of ESC 2i STARR-seq showed a high correlation before and after transfection ($R = 0.95$; Fig. 2.4B). Nevertheless, a majority of NANOG consensus peaks exhibited a higher signal in the STARR-seq sample compared with the vector library, confirming the purifying effect of self-transcription. In contrast, comparison of STARR-seq samples from independent transfections revealed high reproducibility between replicates ($R = 0.95$; Fig. 2.4C). Clustering of ChIP-seq and STARR-seq datasets based on pair-wise correlations supported high reproducibility between STARR-seq replicates which clustered together according to ESC conditions (2i/serum; Fig. 2.4D). In summary, the quality control analysis showed that our STARR-seq data exhibited the necessary fragment complexity as well as reproducibility between biological replicates to be used for downstream analysis.

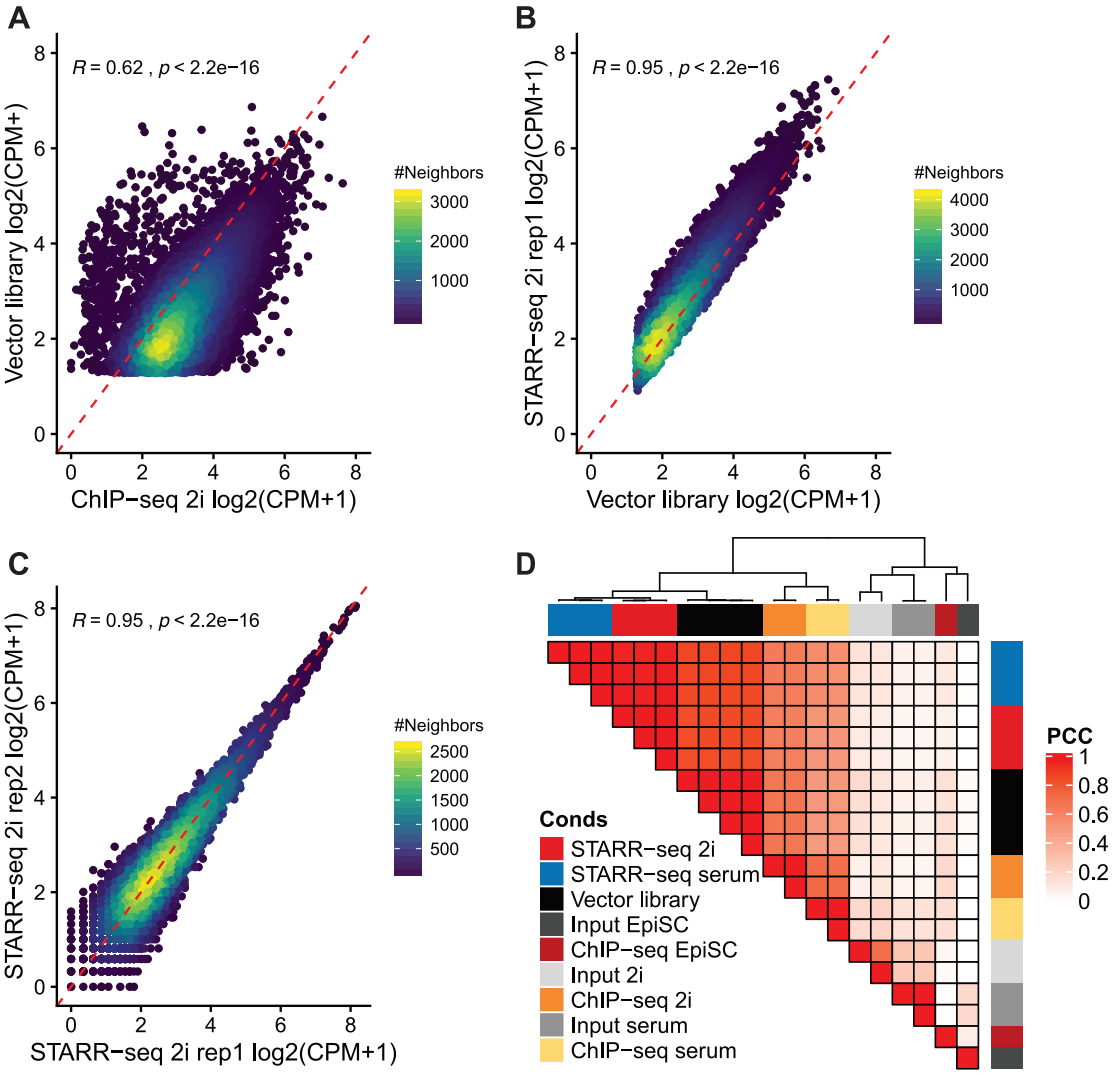


Fig. 2.5: Comparative analysis between ChIP-seq, STARR-seq vector library and STARR-seq samples. **(A)** Scatterplot comparing ESC 2i NANOG ChIP-seq signal against STARR-seq vector library coverage within NANOG consensus peaks. Pearson correlation coefficient and p-value from correlation test is shown. Colour gradient encodes point density. **(B)** Scatterplot comparing STARR-seq vector library coverage and STARR-seq signal within NANOG consensus peaks. Pearson correlation coefficient and p-value from correlation test is shown. Colour gradient encodes point density. **(C)** Scatterplot comparing STARR-seq signals in two different replicates within NANOG consensus peaks. Pearson correlation coefficient and p-value from correlation test is shown. Colour gradient encodes point density. **(D)** Pearson correlation coefficient (PCC) heatmap of vector library, STARR-seq, NANOG ChIP-seq and input samples. Colour annotation encodes for sample condition. Hierarchical clustering of samples is based on 1-PCC as distance measure and ward.D2 as linkage method.

2.3.3 Defining *bona fide* NANOG binding enhancers by comparison of vector library and STARR-seq

In order to determine which NANOG binding sites correspond to *bona fide* transcriptional enhancers, we performed a differential analysis between STARR-seq 2i and associated vector libraries using edgeR [195]. We classified our 13,647 Nanog peaks into three categories based on the differential count analysis: 3,510 active enhancers ($FDR \leq 0.05$ & $log2(FC) > 0$), 2,838 inactive TF binding sites (TFBS; $FDR \leq 0.05$ & $log2(FC) < 0$) and 7,299 grey zone peaks ($FDR > 0.05$). The large number of statistically non-significant regions can be explained by two limitations of our assay. First, a recent publication describing genome-wide mammalian STARR-seq showed that transcription can initiate not only happens at the minimal core promoter sequence but also at the bacterial origin of replication [176]. This reduces the signal-to-noise ratio of STARR-seq and therefore the power to identify *bona fide* enhancer elements. Second, we removed duplicated fragments, since without UMIs these are indistinguishable from PCR amplified fragments [219].

However, inspection of statically significant active enhancers and annotation with nearby genes confirmed their association with known pluripotency genes such as *Esrrb*, *Nanog*, *Klf4*, *Pou5f1* and *Prdm14* (Fig. 2.6A) [220, 183]. In contrast, examination of inactive NANOG TFBS revealed their proximity to genes involved in exit from pluripotency, such as *Foxd3*, and development, such as *Hoxa1* (Fig. 2.6A) [221, 222].

Most active enhancers and inactive TFBS are located distally from genes or within introns, though a higher proportion of active enhancers are distal intergenic compared with inactive TFBS (50.8% (1,783/3,510) compared with 39.5% (1,120/2,838; Fig. 2.6B). In contrast, 25.3% (717/2,838) of inactive TFBS overlap with annotated promoters compared with 14% (4,93/3,510) of active enhancers.

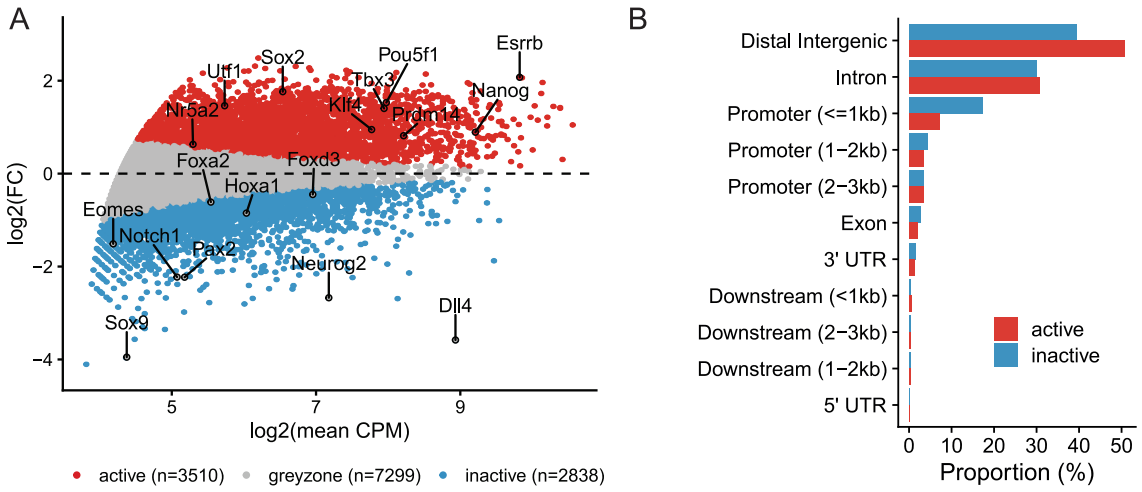


Fig. 2.6: Differential analysis between vector library and STARR-seq to define active enhancers. (A) MA-plot comparison of NANOG consensus peak coverage in STARR-seq 2i against vector library. Peaks were considered active enhancers if $FDR \leq 0.05$ & $log2(FC) > 0$ (red) or inactive TFBS if $FDR \leq 0.05$ & $log2(FC) < 0$. Selected nearby genes involved in ESC pluripotency or development are highlighted. **(B)** Distribution of genome annotations for active enhancers (red) and inactive NANOG TFBS (blue).

As mentioned above, we observed many active enhancers in proximity to well described pluripotency genes and therefore we asked if we would recover previously published enhancer elements (Fig. 2.6A). For example, we recovered a known enhancer element near the *Nanog* gene [223] (Fig. 2.7A). Furthermore, this enhancer exhibited known hallmarks such as binding of NANOG/SOX2/OCT4 binding, chromatin accessibility and enrichment of the histone modifications H3K4me1 as well as H3K27ac. On the other hand, the existence of inactive TFBS near known developmental genes raised questions about whether these sites were previously described as active enhancers during ESC differentiation or *in vitro* development (Fig. 2.6A). A publication by De Kumar *et al* described two NANOG binding sites in the vicinity of *Hoxa1* which "switch on" as an active enhancer upon induction of differentiation by retinoic acid (RA) treatment [222]. We observed this developmental enhancer among our inactive NANOG TFBS (Fig. 2.7B). The binding site displayed binding of NANOG/OCT4/SOX2, low levels of ATAC-seq accessibility but similar levels of H3K4me1. In contrast to the Nanog associated active enhancer above, this site showed H3K27ac depletion but enrichment of the repressive Polycomb mark H3K27me3.

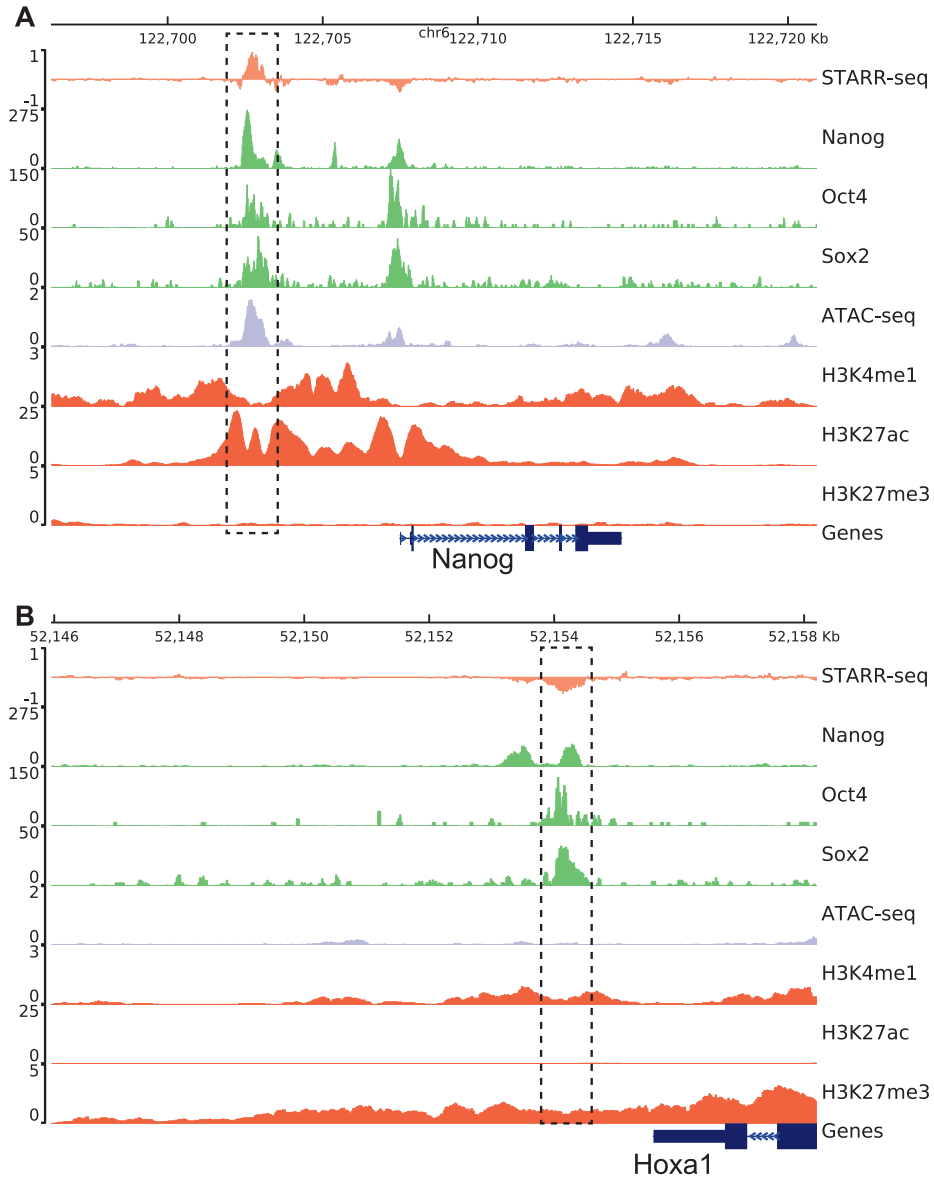


Fig. 2.7: Examples for an active NANOG binding enhancer and inactive NANOG binding site. (A) Genome browser screenshot for a previously described active NANOG binding enhancer upstream of the *Nanog* gene [223]. Signal tracks shown include from top to bottom: STARR-seq log2(FC) (orange); TF ChIP-seq for NANOG, OCT4 and SOX2 as background subtracted signals (green); ATAC-seq signal as CPMs (purple); histone PTM ChIP-seq signal for H3K4me1, H3K27ac and H3K27me3 normalised to CPM (red). The bar along the top denotes genomic position along chromosome 6. **(B)** Genome browser screenshot for a previously described inactive NANOG binding site upstream of *Hoxa1* [222]. Signal tracks as in (A).

Taken together, this analysis established our definition of *bona fide* enhancers and inactive NANOG TFBS from ChIP STARR-seq data. Active enhancers showed the expected ability to

drive transcription from a minimal promoter. In contrast, inactive TFBS were highly engaged by NANOG but did not function as active enhancers in the ESC ground-state (2i) conditions. Moreover, we could find supporting evidence for both classes of NANOG binding sites in the literature.

2.3.4 Transcription factor binding at active enhancers and inactive TFBS

Having classified NANOG binding sites into active enhancers and inactive TFBS, we asked if these sites displayed differential binding of core-pluripotency factors including NANOG, OCT4 and SOX2 [183]. As expected, NANOG showed the smallest average difference in binding signals between active and inactive sites (Fig. 2.8B). However, comparison of Nanog signal distributions over all sites revealed significantly higher ChIP-seq enrichment within active enhancers ($median_{active} = 2.9$, $median_{inactive} = 2.54$, $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test). The STARR-seq and Nanog ChIP-seq signals showed no correlation among active enhancers ($R = -0.07$) and only weak correlation at inactive binding sites ($R = 0.26$) (Fig. 2.8A, B). Both OCT4 and SOX2 display substantial binding to active and inactive sites (Fig. 2.8C, D). However, the TFs show greater occupancies at active enhancers (OCT4: $median_{active} = 0.15$, $median_{inactive} = -0.16$; SOX2: $median_{active} = 0.58$, $median_{inactive} = 0.19$; both: $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test). Neither, TF signal is correlated with the STARR-seq signal within active enhancers (SOX2: $R = 0.02$; OCT4: $R = 0.05$) and only weakly correlated among inactive TFBS (SOX2: $R = 0.17$, OCT4 $R = 0.14$; Fig. 2.8A, C, D).

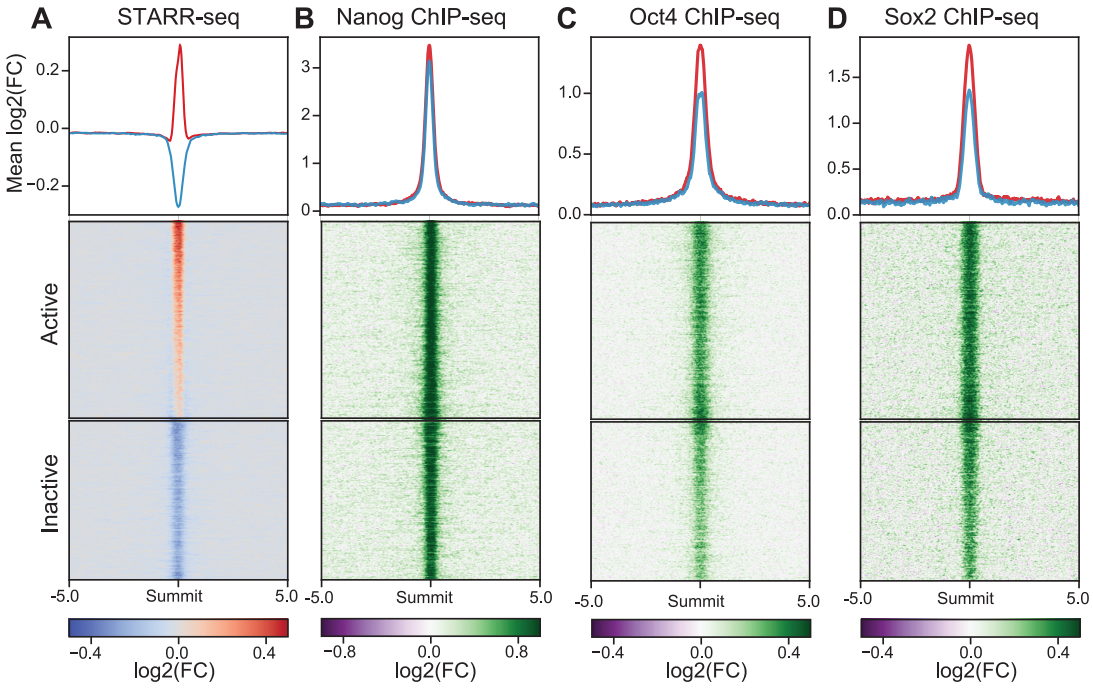


Fig. 2.8: NANOG, OCT4 and SOX2 binding, measured by ChIP-seq, at active enhancers and inactive TFBS. Heatmaps visualise **(A)** STARR-seq signal as log2(FC) over vector library and **(B)** NANOG, **(C)** OCT4, **(D)** SOX2 ChIP-seq signal as log2(FC) over input for $\pm 5kb$ around active enhancer (red metaprofile) and inactive TFBS (blue metaprofile) region summits. Upper panels of the heatmap corresponds to active enhancers and lower panels to inactive NANOG TFBS. Rows are ordered according to STARR-seq signal.

The ChIP-seq assay itself comes with its own obstacles and biases; for example formaldehyde cross-linking might result in false-positive binding sites [218]. To rule out these artefacts in our observations, we sought to confirm independently TF binding to these sites by profiling methods that do not rely on formaldehyde cross-linking, such as CUT&RUN [119]. We made use of a recently published CUT&RUN dataset for NANOG, OCT4 and SOX2 [120]. This analysis confirmed that all three TFs bind to both classes of NANOG peaks (Fig. 2.9A). Further, all three TFs showed significantly higher engagement to active enhancers rather than inactive binding sites (NANOG: $median_{active} = 0.99$, $median_{inactive} = 0.8$, $p < 2.2 \cdot 10^{-16}$; OCT4: $median_{active} = 0.58$, $median_{inactive} = 0.58$, $p = 5.4 \cdot 10^{-14}$; SOX2: $median_{active} = 0.8$, $median_{inactive} = 0.52$, $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.9B). Comparison with CUT&RUN data further evidenced that observed STARR-seq signal was not correlated to TF binding signals in active enhancers (NANOG: $R = 0.06$, OCT4: $R = 0.04$, SOX2: $R = 0.08$) nor in inactive TFBS (NANOG: $R = 0.09$, OCT4: $R = 0.05$, SOX2: $R = 0.11$). Although this analysis confirmed binding of NANOG, OCT4 and SOX2 to both classes

of NANOG peaks, it also supported weaker binding of these TFs to inactive TFBS. These observations lead to a fundamental question of whether there is less TF engagement to inactive TFBS within ESC populations.

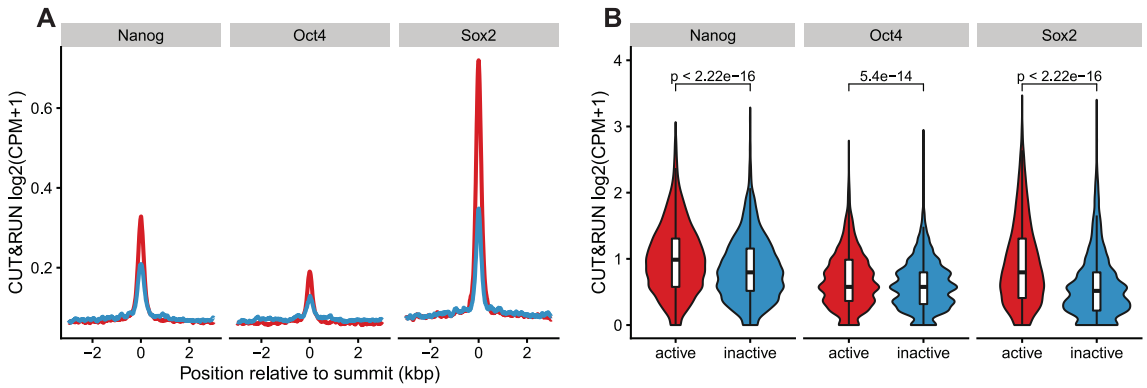


Fig. 2.9: CUT&RUN signal for NANOG, OCT4 and SOX2 at active enhancers and inactive TFBS. (A) Metaprofiles visualising average CUT&RUN binding signal of NANOG, OCT4 and SOX2 at active enhancers (red) and inactive binding sites (blue). **(B)** Violin plot comparison of CUT&RUN NANOG, OCT4 and SOX2 signals within active enhancer and inactive binding sites (summed within ±250bp around summit). P-values from one-sided Wilcoxon test.

To answer this question, we utilised ATAC-seq a technique to profile open-chromatin by sequencing adapter insertion with a hyperactive Tn5 transposase [131, 224]. This method results in fractions of different fragment lengths representing either occupancy of TF or other chromatin associated proteins ($< 140bp$), mono-nucleosomes ($> 141bp$ & $< 320bp$) or longer nucleosome arrays ($> 320bp$). Our initial observation was that both NANOG peak classes exhibited significant enrichment of TF fragments compared with genomic background ($p < 2.2 \cdot 10^{-16}$, Binomial test). In contrast, only inactive binding sites showed a significant enrichment for nucleosome fragments (inactive: $p < 2.2 \cdot 10^{-16}$, active: $p = 0.14$, Binomial test). Comparison of fragment length distributions revealed a higher proportion of TF fragments overlapping active enhancers and more mono-nucleosomes fragments within inactive binding sites (Fig. 2.10A). Nevertheless, most NANOG peaks, independent of their activity, showed more TF fragments than expected but active enhancers were significantly higher enriched than inactive binding sites ($p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.10B *right panel*). This was further supported by a well-defined TF footprint around the OCT4-SOX2-TCF-NANOG composite motif detectable in both NANOG peak classes (Fig. 2.10D). This observation provided further support for NANOG engagement to inactive TFBS. In contrast, most inactive binding sites showed more mono-nucleosome fragments than expected

($p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.10B *left panel*). Active enhancers showed on average not more mono-nucleosome fragments than expected.

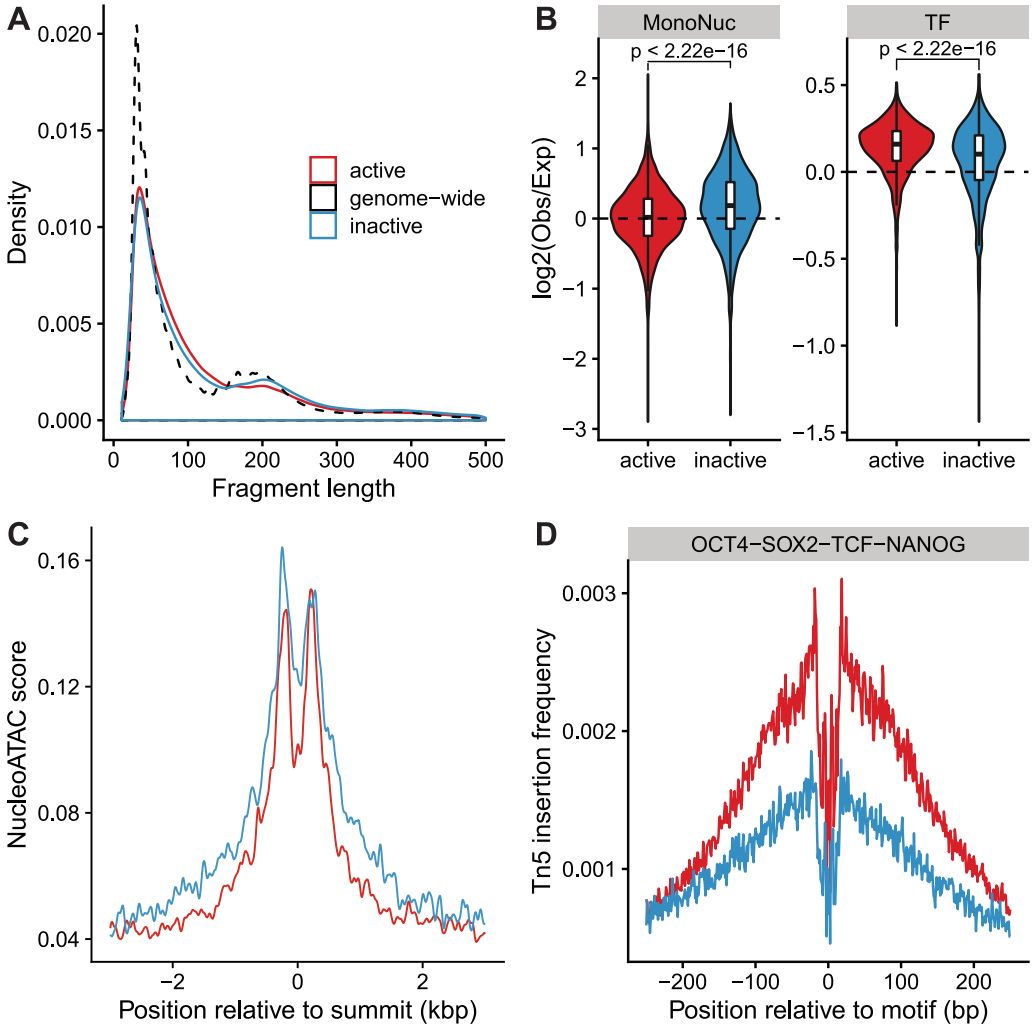


Fig. 2.10: TF and nucleosome occupancy as measured by ATAC-seq in active enhancers and inactive binding sites. (A) ATAC-seq fragment length distribution stratified for all fragments (genome-wide, black), active enhancers (red) and inactive binding sites (blue). **(B)** Enrichment of TF fragments (< 140) and nucleosome fragments (> 140) as log2 observed over expected for each peak. P-value from two-sided Wilcoxon test. **(C)** Average NucleoATAC nucleosome occupancy scores for $\pm 2.5kb$ around peak summits for either active enhancers (red) or inactive binding sites (blue). **(D)** TF footprint of Tn5 insertions around NANOG-SOX2-OCT4 motif matches from either active enhancers (red) and inactive binding sites (blue).

To follow up on the depletion of nucleosome fragments from active enhancers we performed nucleosome calling using NucleoATAC [136]. The extracted nucleosome signal revealed for both NANOG peak classes two flanking mono-nucleosomes on average but higher nucleosome depletion within peaks in active enhancers (Fig. 2.10C). This was further supported by

the significant enrichment of nucleosome free regions (NFR; determined with NucleoATAC) overlapping active enhancers compared with inactive binding sites ($p = 7.46 \cdot 10^{-5}$, one-sided Fisher's exact test). Our ATAC-seq analysis confirmed that inactive TFBS exhibited a higher nucleosome density and less TF binding compared with active enhancers under ESC ground-state (2i) conditions.

The previous observations of less TF engagement at inactive binding sites led us to ask if there are other TFs which might co-bind with Nanog or if we could find evidence that these sites were occupied by any other TFs in differentiated cells. To answer these questions we performed enrichment analysis of TF ChIP-seq peaks using two different databases called LOLA (545 TF ChIP-seq samples) and Cistrome (2,381 TF ChIP-seq samples; Fig. 2.11) [192, 193]. Results are presented as volcano-like plots in which the x-axis corresponds to log2(Odds) from a one-sided Fisher's exact test as measure of enrichment and the y-axis displays the Matthews Correlation Coefficient (MCC) as measure of predictive power computed from the same contingency table (see Methods; Fig. 2.11A, B). At first glance this analysis confirmed our previous observations regarding TF binding at active enhancers. We observed higher enrichment and predictive power for TFs that are already described in the literature for their association with pluripotency including SOX2, ESRRB, POU5F1, NR5A2, NR0B1 and SMAD3 [220, 225]. Even NANOG peaks from these public datasets showed higher enrichment for our set of active enhancers, suggesting a dependence on the peak calling procedure (Fig. 2.11A). Interestingly, OTX2 and RARA peaks exhibited higher enrichment within active enhancers, although both TFs were more involved in ESC differentiation [226, 227]. In contrast, inactive binding sites showed only enrichment of ZFP57 peaks from ChIP-seq in ESCs, a factor which was previously described as being involved in imprinting and maintenance of repressive histone marks [228]. Moreover, we observed enrichment of various factors best known for their functions in differentiated tissues including TBX5, TBX6, GATA2, NEUROD2, ERG, RUNX1 and MYC (Fig. 2.11A, B). Binding of these regulators in differentiated cells might point towards a role of these inactive TFBS beyond the pluripotent cell state. For example, a recent publication suggested that pluripotency factors might serve as bookmarking factors at regulatory elements which become active during differentiation [229].

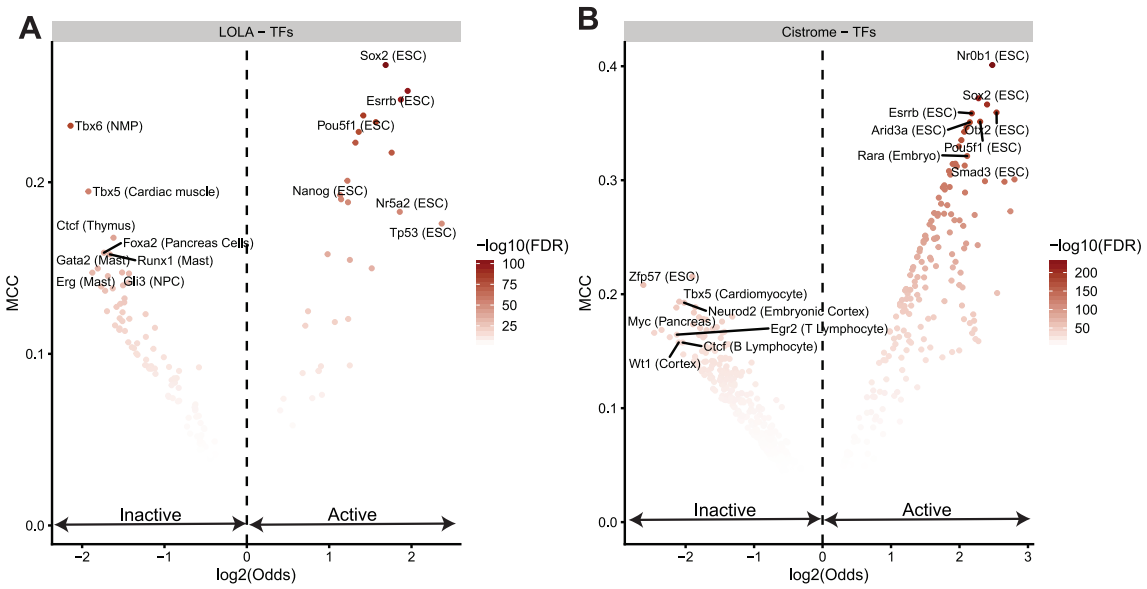


Fig. 2.11: Differential enrichment analysis of TFs between active enhancers and inactive binding sites. Volcano-like plot visualisation of ChIP-seq peak enrichment analysis results for known TFs [63]. ChIP-seq peaks were obtained from **(A)** LOLA **(B)** Cistrome databases [192, 193]. Log-odds values from Fisher's exact tests were plotted against Matthews correlation coefficient (MCC). Colour gradient encodes $-log10(FDR)$ derived from Fisher's exact test followed by Benjamini-Hochberg correction. Top 8 most enriched and predictive features were labelled.

Taken together, we showed that active enhancers exhibited all TF related features that represent *bona fide* enhancer elements, such as engagement by various TFs, including key pluripotency regulators and nucleosomes depletion. At inactive sites, we also observed NANOG binding, but in contrast found high nucleosome densities as well as revealing potential TF binding.

2.3.5 Chromatin features of active enhancers and inactive TFBS

Another important hallmark of enhancers is the chromatin signature associated with these regulatory elements [83, 230]. It is established that the histone marks H3K4me1 without H3K27ac identifies so called primed enhancers, while the presence of H3K27ac marks active enhancers [230, 161]. We asked whether we recovered these histone PTM signatures in our active enhancers and what chromatin context our inactive TFBS showed. First, we examined chromatin accessibility within both Nanog binding site classes, again using published ESC 2i ATAC-seq data [224]. Additionally, to previous observations of differential TF/nucleosome engagement, this revealed that active enhancer regions exhibited significantly higher accessi-

bility compared with inactive binding sites ($median_{active} = 2.96$, $median_{inactive} = 2.37$, $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test). Nevertheless, independent of their activity all Nanog peaks showed enrichment of ATAC-seq fragments and therefore accessibility (Fig. 2.12B). Overall no correlation between chromatin accessibility and STARR-seq signal was observable for both peak sets (active: $R = 0.14$, inactive: $R = -0.02$). Next, we examined published ChIP-seq datasets for H3K4me1, H3K27ac and the repressive mark H3K27me3 [231]. The primed enhancer mark H3K4me1 was observed at active enhancers and inactive TFBS in nearly indistinguishable amounts ($median_{active} = 1.03$, $median_{inactive} = 1.02$, $p = 0.92$, two-sided Wilcoxon test; Fig. 2.12C). As expected, H3K27ac showed higher levels at active enhancers compared with inactive TFBS ($median_{active} = 2.27$, $median_{inactive} = 0.714$, $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.12D). In contrast, H3K27me3 showed much higher levels at inactive TFBS ($median_{active} = 0.19$, $median_{inactive} = 0.31$, $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.12E). Interestingly for all three histone marks a clear NFR was apparent around the centre of NANOG binding sites (Fig. 2.12C, D, E). We did not observe a correlation between STARR-seq signal and one of the three histone marks neither within active enhancers (H3K4me1 $R = 0.11$, H3K27ac $R = 0.18$, H3K27me3 $R = -0.12$) nor for inactive TFBS (H3K4me1 $R = 0.023$, H3K27ac $R = 0.05$, H3K27me3 $R = -0.11$).

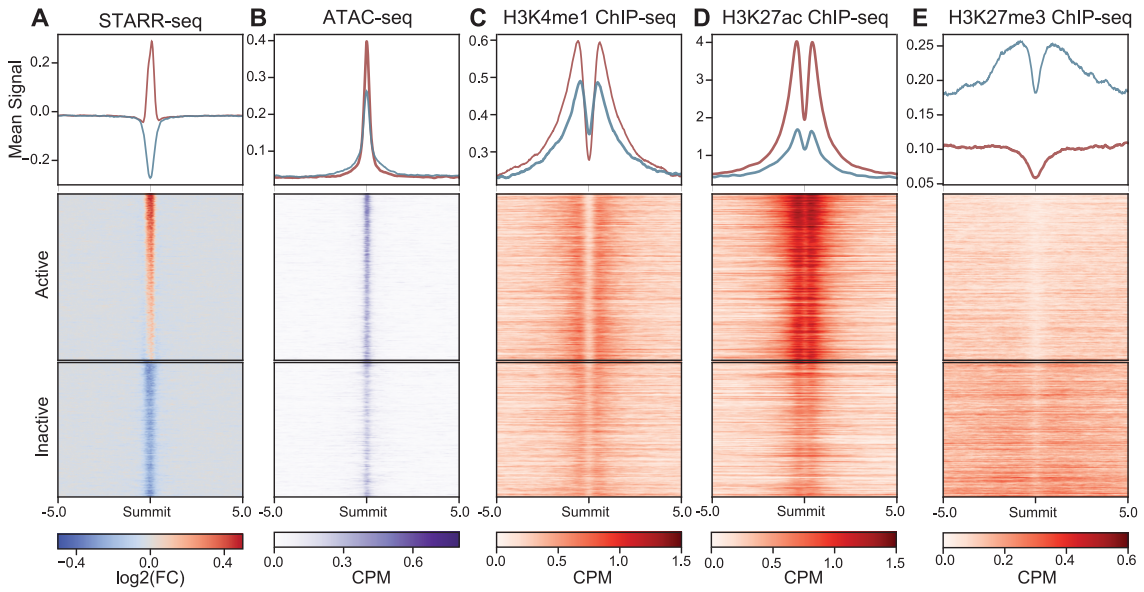


Fig. 2.12: Histone PTM ChIP-seq and ATAC-seq signal around active enhancers and inactive TFBS. Heatmaps show **(A)** STARR-seq signal as log2(FC) over vector library, **(B)** ATAC-seq signal as CPM and **(C)** H3K4me1, **(D)** H3K27ac, **(E)** H3K27me3 ChIP-seq signal as counts per million (CPM) for $\pm 5kb$ around active enhancers (red metaprofile) and inactive binding sites (blue metaprofile) region summits. Upper panels of the heatmap corresponds to active enhancers and lower panels to inactive TFBS. Rows are ordered according to STARR-seq signal.

Due to observable differences in the chromatin landscape of active enhancers and inactive TFBS, we asked if there might be other histone modifications that differentiate them. In order to examine this, we performed differential enrichment analysis using Fisher's exact test in combination with histone modification peaks from the Cistrome database [193]. The database includes 5,005 ChIP-seq samples for various histone modifications and cell types, including 1,093 ESC samples for 35 different histone marks. This analysis revealed differential enrichment of H3K27ac ChIP-seq peaks, from ESC as well as iPSC, for our active enhancer set and thereby confirmed our observation above (Fig. 2.13, 2.12D). Moreover, H3K9ac, another activating histone acetylation, as well as H3.3, which was previously associated with active enhancer elements showed enrichment in our active enhancer set [232, 230].

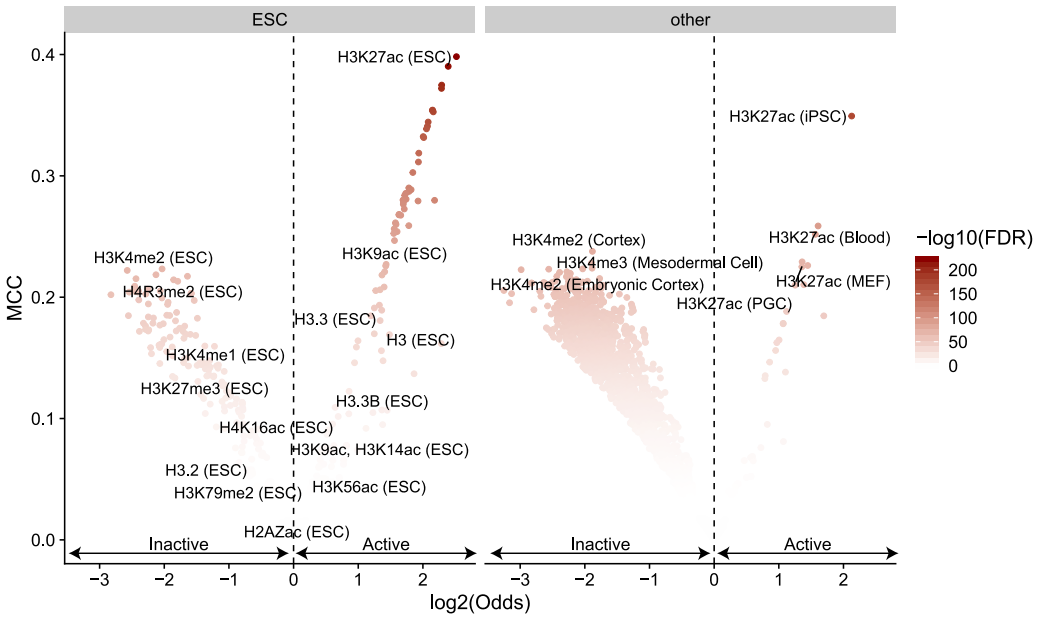


Fig. 2.13: Differential enrichment analysis of histone modifications between active enhancers and inactive TFBS. Volcano-like plot visualisation of histone modification ChIP-seq peak enrichment analysis results. Histone modification ChIP-seq peaks were obtained from the Cistrome database [193]. Log-odds values from Fisher's exact tests were plotted against Matthews correlation coefficient (MCC). Colour scale encoded $-log10(FDR)$ derived from Fisher's exact test followed by Benjamini-Hochberg correction. Top 8 most enriched/predictive features were labelled for mESC and top 3 features were labelled for other cell types.

On the other hand, inactive TFBS exhibited enrichment of H3K4me1 and H3K4me2 in ESC but also in more differentiated tissues and cell types (Fig. 2.13). Both histone modifications were previously described as histone modifications marking primed enhancers [230]. The second most highly enriched histone modification for the inactive TFBS was H4R3me2,

which was previously characterised as a repressive mark as well as being involved in recruiting DNTM3A to establish DNA methylation [233]. Finally, this analysis provided additional support for the enrichment of repressive H3K27me3 at inactive TFBS (Fig. 2.13).

Histone marks are established by so called epigenetic writers, which are chromatin associated proteins that can covalently modify (e.g. methylate or acetylate) histone side chains [234]. Therefore, we hypothesised that the observable differences in H3K27me3 and H3K27ac should be accompanied by similar differences in binding of associated proteins. To test this we again performed ChIP-seq peak enrichment analysis using LOLA and Cistrome with access to 88 and 473 ChIP-seq samples respectively for chromatin associated proteins (Fig. 2.14) [192, 193]. We observed enrichment of the histone acetylase EP300 and its associated co-activator CREBBP which are both crucial for H3K27ac modification [235]. EP300 was previously suggested as predictor for enhancer activity [236]. Moreover, various other chromatin binding proteins previously linked to enhancer activity were found to be enriched in our active enhancer set, including: ASH2L, BRD4, CDK8, CDK9, HDAC1, HDAC2, KDM1A and MED1 [237, 238, 239, 240, 241, 56]. For the inactive binding sites, we observed enrichments of EZH2 and SUZ12 which are part of the Polycomb repressive complex (PRC2) as well as JARID2 and ATRX which both direct PRC2 function to its target sites [242, 243]. Besides these PRC2 associated proteins, we also detected enrichment of KDM2B and UTF1. KDM2B function was previously described as a repressor of developmental genes through recruitment of PRC1 [244, 245]. UTF1 was previously associated with transcriptional regulation from bivalent chromatin [246].

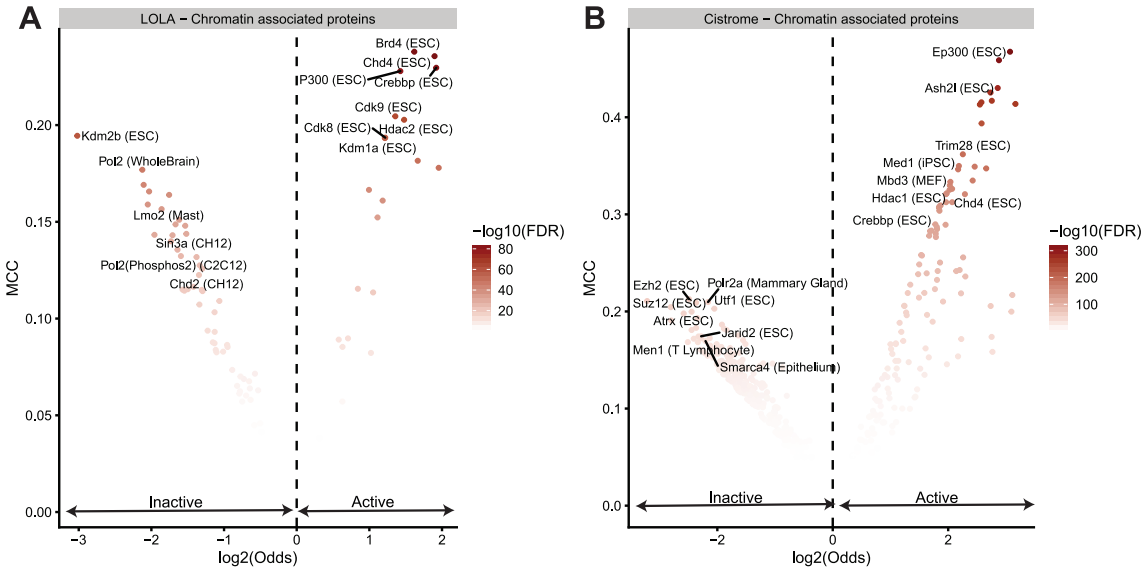


Fig. 2.14: Differential enrichment analysis of chromatin associated proteins between active enhancers and inactive TFBS. Volcano-like plot visualisation of ChIP-seq peak enrichment analysis results for chromatin associated proteins (excluding TFs). ChIP-seq peaks were obtained from **(A)** LOLA and **(B)** Cistrome database [192, 193]. Log-odds values from Fisher's exact tests were plotted against Matthews correlation coefficient (MCC). Colour scale encoded $-log10(FDR)$ derived from Fisher's exact test followed by Benjamini-Hochberg correction. Top 8 most enriched and predictive features were labelled.

Taken together, the data revealed differences in the chromatin landscape associated with active enhancers compared with inactive TFBS. NANOG binding active enhancers were characterised by higher levels of chromatin accessibility as well as H3K27ac enrichment. In contrast, inactive NANOG binding sites showed lower levels of accessibility and were enriched for the PRC2 repressive mark H3K27me3. In both cases we validated these observations using ChIP-seq data from database collections and found further support in the specific enrichment of chromatin associated proteins.

2.3.6 Nascent transcription of ncRNA from active enhancers and inactive TFBS

Yet another known hallmark of active transcriptional enhancers is the transcriptional output of ncRNAs called enhancer RNAs (eRNAs) [161]. In order to quantify eRNAs we made use of a previously published GRO-seq dataset [247]. GRO-seq is a technique which profiles RNA from nascent transcription by bromouridine labelling [159]. We observed bi-directional nascent transcription for our active enhancers (Fig. 2.15A). Inactive binding sites also showed

some bi-directional expression pattern but to a lesser extent than active enhancers. This was in line with previous literature describing this as an important feature of enhancers [36]. Active enhancers exhibited statistically significantly higher nascent eRNA transcription compared with inactive TFBS ($p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.15B). However, there was no correlation between eRNA abundance and STARR-seq signal (active: $R = 0.11$, inactive $R = -0.01$). In summary, in this section we confirmed that our active enhancers recover the previously described feature of bi-directional eRNA expression.

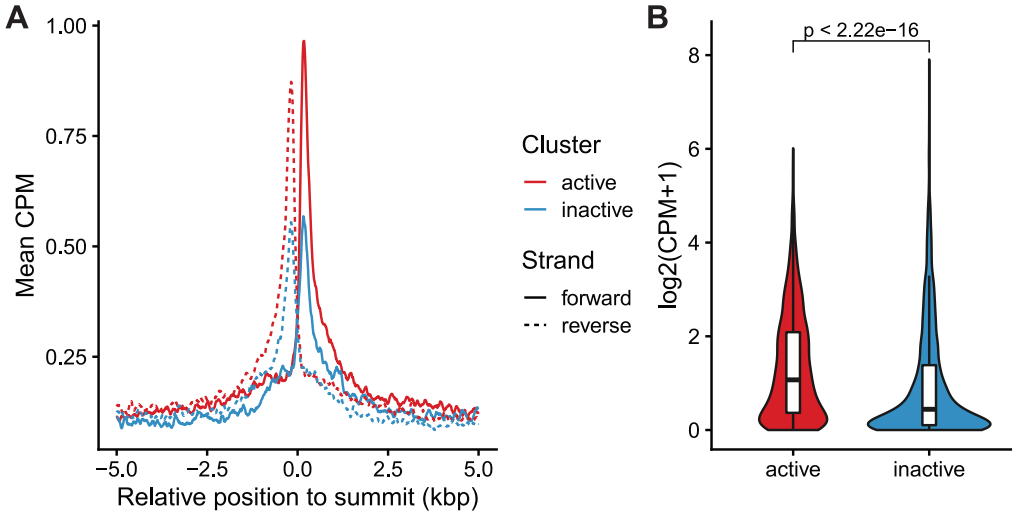


Fig. 2.15: Nascent transcription of ncRNAs at active enhancers and inactive binding sites. (A) Metaprofile summarises strand-specific nascent transcription, measured by GRO-seq, $\pm 5kb$ around the summit of active enhancers (red) and inactive binding sites (blue). Line type indicates strand. GRO-seq signal was normalised to counts per million (CPM). **(B)** Violin plot shows sum of signal within a $\pm$250bp window around the summit of active enhancers and inactive TFBS. P-value from one-sided Wilcoxon test.

2.3.7 Chromatin-chromatin interactions of active enhancers and inactive TFBS

It has been suggested in that transcriptional enhancers influence gene expression by physical interaction with their target promoter in a spatial context [42]. Therefore, we asked if active enhancers and inactive TFBS would show observable differences in their three-dimensional interactions. To address this, we utilised chromatin-chromatin interactions from a previously published capture DHS Hi-C dataset [145]. Capture Hi-C, like all other 3C methods, first digests cross-linked DNA with a restriction enzyme and then ligates the ends of fragments which were in close proximity to each other [248]. This gives the experiment a resolu-

tion limit that depends on the employed restriction enzyme and the density of restriction sites in the profiled genome. After stringent filtering (see Methods), we ended up with 472,961 significant interactions between 113,111 genomic fragments defined by DpnII digests. Mapping our NANOG peaks to these digests revealed that 85.9% (3,013/3,508) of active enhancers and 84.8% (2,407/2,838) of inactive TFBS were unambiguously mappable to one of them. To avoid potentially biasing the analysis by multiple assignments, we removed 430 digests overlapping with more than one NANOG peak, which were associated with 3.4% (15,996/472,961) of all chromatin-chromatin interactions. In total, active enhancers and inactive TFBS were involved in 84,193 interactions, which were used in the following downstream analysis. We grouped interactions into three categories depending on whether the other of the contact involved a promoters (P), another NANOG peak, or any other (O) genomic element (Fig. 2.16A). We also grouped interactions depending whether the NANOG peak overlapped with the bait or OE fragment. This revealed that, independent of the category, most interactions are between a NANOG peak and other genomic elements such as insulators (inactive-O OE=29.90%, active-O bait=38.90%; Fig. 2.16A). In contrast, between 9.00%-12.10% of all interactions were with promoter elements and, even more infrequently, between each other. Inactive-inactive interactions were the least observed comprising only 2.00% of interactions. Since active enhancers and inactive TFBS showed similar preferences of interaction partners, we examined if there are any differences in other interaction features such as strength (CHiCAGO score), number of interactions or distance; however, we did not find any significant differences between active enhancers and inactive TFBS (Fig. 2.16B, C, D) [148]. Our initial analysis of interaction types led us to the question if there might be differences in terms of interactions with other genomic regions. To address this, we obtained a recently published high-confidence interaction dataset, which was based on the same capture Hi-C experiments as our analysis, but annotated with various other regulatory elements [199]. Active enhancers exhibited significant enrichment of two interaction types including "Strong enhancer" to "Strong enhancer" as well as "Weak promoter" to "Strong enhancer" (Fig. 2.16E; $FDR = 6.3 \cdot 10^{-47}$, $FDR = 10^{-4}$, two-sided Fisher's exact test followed by Benjamini-Hochberg correction). These observations were in line with previous observed differences in histone PTM ChIP-seq (Fig. 2.12). In contrast, inactive TFBS showed significant enrichment of 15 different interaction types (Fig. 2.16E). Amongst the most abundant of these interactions were "CTCF binding site" to "Strong promoter", "Bivalent promoter" to "CTCF binding site" and "Bivalent region" to "Bivalent promoter" ($FDR = 7.9 \cdot 10^{-15}$, $FDR = 7.9 \cdot 10^{-12}$, $FDR = 5.6 \cdot 10^{-}9$, two-sided Fisher's exact test followed by Benjamini-Hochberg correction). These observations are in line with our previous

analysis showing H3K27me3 enrichment as well as differential enrichment of CTCF at our inactive binding sites (Fig. 2.11 & Fig. 2.12). In summary, no significant differences were observable for basic interactions features such as interaction strength, number or distance between active enhancers and inactive enhancers. However, we found that the chromatin state of interacting regions was specific to active and to inactive sites and were in line with our previous observations of their respective chromatin landscape.

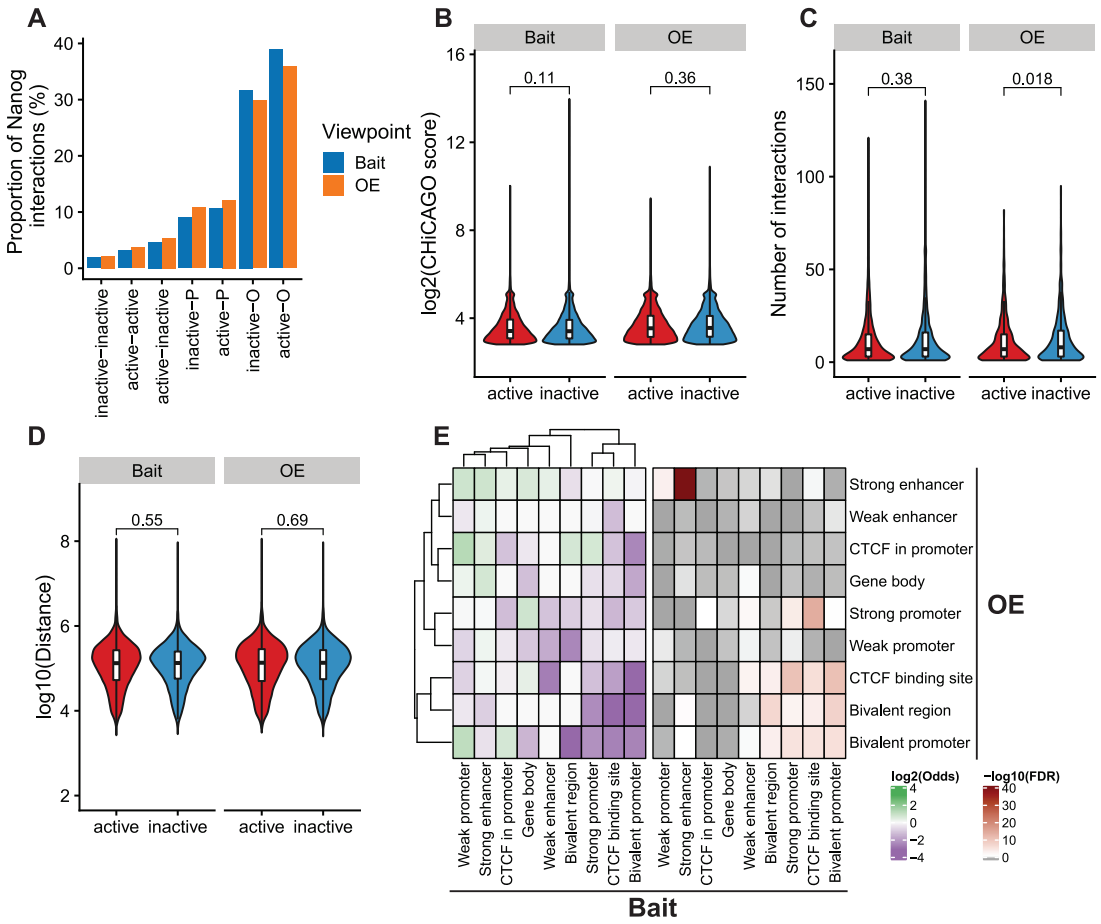


Fig. 2.16: Chromatin-chromatin interactions of active enhancers and inactive binding sites. **(A)** Summary of interaction types involving active enhancers and inactive TFBS including amongst themselves, with promoters (P) or other (O) genomic elements. Colours encode if NANOG peak was overlapping digest which was baited (blue) or "other ends" (OE) (orange). **(B)** Comparison of interaction strengths as measured by CHiCAGO scores [148]. **(C)** Comparison of the number of significant interactions made by a peak. **(D)** Comparison of interaction distances between baited fragments (left panel) and OE fragments (right panel). Colour encodes if the respective fragment was overlapping with active enhancer (red) or with an inactive binding site (blue). P-values from two-sided Wilcoxon test. **(E)** Differential Chromatin state enrichment analysis between interactions outgoing from active enhancers compared with inactive binding sites. Heatmaps visualising log-odds (*left*) and Benjamini-Hochberg corrected p-value (*right*) from Fisher's exact test.

2.3.8 Influence of active enhancers and inactive TFBS on gene expression

The most important feature of transcriptional enhancers is their ability to enhance transcriptional output of their target gene [42]. Therefore, we investigated if there were observable differences in target gene expression of active enhancers and inactive TFBS. To examine these relationships robustly, we performed the following analysis based on three different approaches of assigning NANOG peaks to genes (see Methods). First, we assigned peaks to target genes based on published DHS capture Hi-C data from ground-state (2i) ESC [145]. Target genes were grouped into three categories depending whether they interact only with active enhancers ('active'), with both types ('both') or only with inactive TFBS ('inactive'). Using the DHS capture Hi-C, we obtained 1,592 active, 920 both and 1,576 inactive target genes. Comparing their RNA-seq expression levels revealed significant differences between all three of them (Fig. 2.17A). Genes interacting with both types of peaks showed highest median expression, whereas genes only interacting with inactive TFBS showed the lowest expression. However, the DHS capture Hi-C data we used had a mean resolution of 3,676bp which made the definitive assignment of peak-gene pairs impossible. Therefore, we next used a previously published high-resolution capture Hi-C dataset resolving interactions of regions below 1,000bp [146]. This dataset was generated under primed (serum) ESC conditions. However, a recently published study suggested only marginal changes in enhancer promoter interactions between 2i and serum ESC [199]. Using these high-resolution interactions, we could assign 465 active, 50 both and 273 inactive target genes. Again, the three target gene sets showed significant differences between each other (Fig. 2.17B). Genes interacting with both peak types exhibited the highest median expression while those only interacting with inactive TFBS showed the lowest expression. Finally, we performed a similar analysis using GREAT, a tool to assign nearby genes to peaks [128]. This approach defined 2,688 active, 1,588 both and 2,002 inactive target genes. Comparison of RNA-seq expression revealed only significant differences between active/both and inactive/both (Fig. 2.17C). Genes assigned to active enhancers and inactive binding sites appearing equally low expressed.

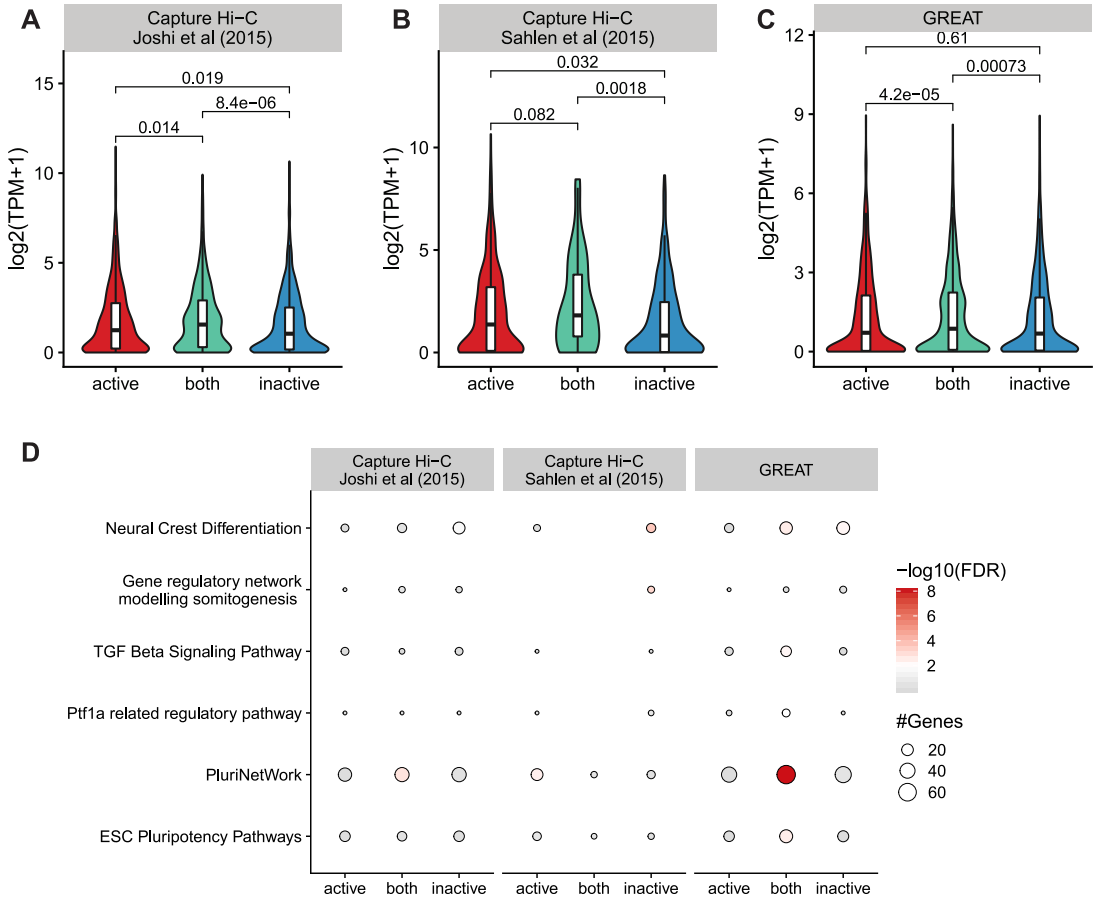


Fig. 2.17: Expression of active enhancers and inactive TFBS target genes. Violin plots showing gene expression as measured by RNA-seq for three different target gene sets based on **(A)** DHS capture Hi-C from Joshi *et al*, **(B)** high resolution capture Hi-C from Sahlen *et al* and **(C)** nearby gene association with GREAT. Genes were stratified into those only interacting with active enhancers ('active', red), active as well as inactive sites ('both', green) and inactive TFBS only ('inactive', blue). RNA-seq counts were normalised to transcripts per million (TPM). P-values from two-sided Wilcoxon test. **(D)** Pathway enrichment for the three different target gene sets and stratified according to interaction preference. Dot size corresponds to number of genes overlapping with pathway. Colour gradient encodes $-log10F(FDR)$ from one-sided Fisher's exact test followed by Benjamini-Hochberg correction. Pathway information was downloaded from WikiPathway [201].

Next, we investigated the cellular functions of the genes from the three different gene sets by performing a pathway enrichment analysis. This showed significant enrichment of genes involved in the pluripotency network ('PluriNetWork') for all three target-gene assignment approaches (Fig. 2.17D). Although, for the Joshi *et al* capture Hi-C and GREAT approach, genes interacting with both showed significant enrichment ($FDR = 1.53 \cdot 10^{-3}$, $FDR = 9.8 \cdot 10^{-9}$, Benjamini-Hochberg corrected one-sided Fisher's exact test). In contrast,

gene sets based on the Sahlen *et al* capture Hi-C showed enrichment of this pathway in the active gene set ($FDR = 3.46 \cdot 10^{-3}$, Benjamini-Hochberg corrected one-sided Fisher's exact test). Thus, providing additional support for the specificity of our active enhancer set. Interestingly, we observed significant enrichment of the "neural crest differentiation" pathway amongst genes associated with inactive TFBS for all three approaches (Joshi $FDR = 0.02$, Sahlen $FDR = 0.0002$, GREAT $FDR = 0.005$, Benjamini-Hochberg corrected one-sided Fisher's exact test). This observation is in line with previous publications describing NANOG as marker for neural crest stem cells in chick but also as major repressor of neural crest differentiation in human ESC [249, 250].
Taken together, in this section we showed that genes associated with active enhancers or both showed higher expression than genes linked to inactive TFBS. Further we could provide evidence that our active enhancers physically contact key pluripotency genes whereas inactive TFBS interact with genes involved in differentiation/development.

2.3.9 Differences in core-promoter architecture of active enhancers and inactive TFBS

Recent studies revealed an intrinsic specificity between enhancer elements and certain core promoter architectures of their target genes [47, 48]. This means that observable enhancer activities in our ChIP STARR-seq assay might depend on the minimal promoter used in the screening vector. Therefore, we examined whether our set of inactive TFBS were truly inactive or just lacked the ability to enhance transcription in combination with the pGL3 minimal promoter. To rule out such a potential bias, we made use of our three target gene sets to perform a differential core-promoter element enrichment analysis (see Methods). For both capture Hi-C gene sets, we found no any significant differences in terms of core-promoter enrichment (Fig. 2.18A, B). Only the nearby gene set assigned by GREAT showed significant differences in GC-box enrichment ($FDR = 0.01$, Benjamini-Hochberg corrected two-sided Fisher's exact test). Since two out of three gene sets did not show any significant differences in terms of core-promoter enrichment between both classes, we conclude that there is only weak evidence for differential core-promoter usage between the two NANOG binding classes. Moreover, considering our previous observations revealing differences in terms of TF binding and associated chromatin landscapes, we excluded the possibility that these observations could be an artefact of our experimental design.

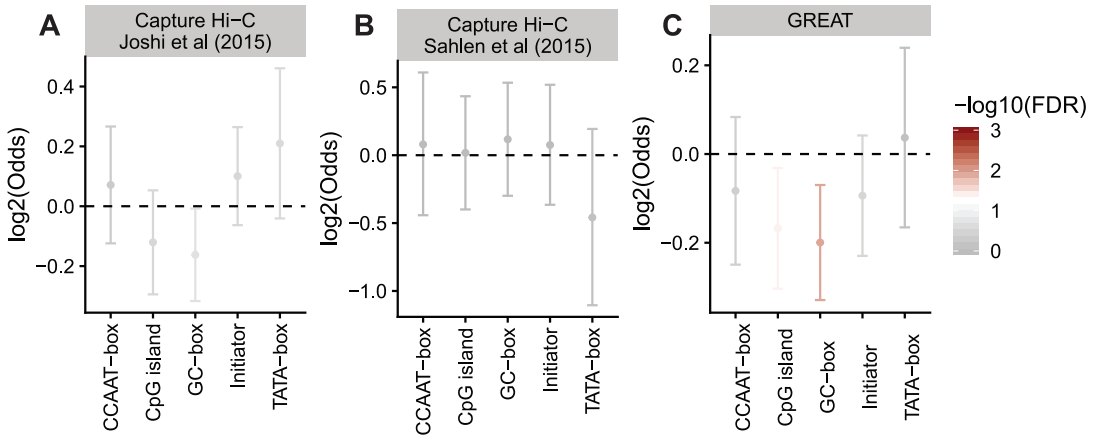


Fig. 2.18: Enrichment of core-promoter elements in target gene sets assigned to active enhancer and inactive TFBS. Differential core-promoter enrichment between gene sets associated with active enhancers and inactive binding sites. Gene sets were derived from **(A)** Joshi et al (2015) capture Hi-C, **(B)** Sahlen *et al* (2015) capture Hi-C and **(C)** assignment of nearby genes using GREAT. Log-odds were computed with a two-sided Fisher's exact test comparing active and inactive genes. P-values were derived from same two-sided Fisher's exact test and Benjamini-Hochberg corrected.

2.3.10 Predicting active enhancers and inactive TFBS from DNA sequence alone

Our prior analysis uncovered substantial differences, in TF binding, local chromatin landscape and putative target gene expression between active enhancers and inactive TFBS. These characteristics influence each other; for example binding of some TFs depend on chromatin accessibility which in turn depends on histone modifications [42, 34]. On the other hand, histone modifications are dependent on the recruitment of co-factors following specific TF binding [42, 34]. Ultimately, all these processes are encoded in the DNA sequence itself to a certain extent via TF binding motifs and other sequence patterns. This means if a particular Nanog binding site corresponds to an active enhancer, information on this is encoded in its underlying DNA sequence. We therefore asked what sequence features were predictive to distinguish active enhancers from inactive TFBS.

To answer this question, we made use of three different machine learning algorithms and trained them on DNA sequence features to classify active enhancers and inactive TFBS. These models include gradient boosted machines (GBM), gapped-kmer support vector machines (gkm-SVM) and convolutional neural networks (CNN) [211, 208, 210]. In the following sections, we will discuss how we defined input features, optimised model hyperparameters and performed model training.

Sequence feature selection, GBM training and hyperparameter selection

We first trained GBMs, a tree-based model, since these are known to have a good trade-off between performance and interpretability [211]. However, before training a GBM it was necessary to extract and select meaningful features from DNA sequences underlying NANOG peaks. To achieve this, we devised a novel two-step strategy for TF selection using previously published proteomics datasets in combination with TF binding predictions using position weight matrix (PWM) scoring. In the first step, we performed a meta-analysis of NANOG co-binding TFs as measured by ChIP-SICAP as well as the chromatin-bound proteome (Chromatome) in ESCs to define biologically meaningful features [204, 205]. This revealed 29 Nanog co-binding TFs which were detectable independent of study and experimental design (2i vs serum compared with ChIP vs no antibody; Fig. 2.19A). Amongst these NANOG co-binding TFs we found well known pluripotency factors such as SOX2, OCT4 and ESRRB [183]. We also defined a second feature set as a control, consisting of the remaining 86 TFs detected in the whole chromatin proteomics but not overlapping with the NANOG ChIP-SICAP dataset, indicating no co-binding with NANOG (Fig. 2.19A). The control set contained ubiquitously expressed factors such as SP1 and CTCF but also ESC specific factors such as ZIC3, KLF2 and TCF3 [251, 252, 253]. Comparison of NANOG co-binding and control TFs revealed differences in their TF family affiliation; for example ARID binding TFs were specific NANOG co-binders while bZIP TFs could only be found among control TFs (Fig. 2.19B). To further ensure the validity of our NANOG co-binding TFs, we performed pathway enrichment analysis using pathway gene sets from WikiPathway [201]. This confirmed the enrichment of both TF sets in the pluripotency network and revealed enrichment of nuclear receptors among the NANOG co-binding TFs (Fig. 2.19C).

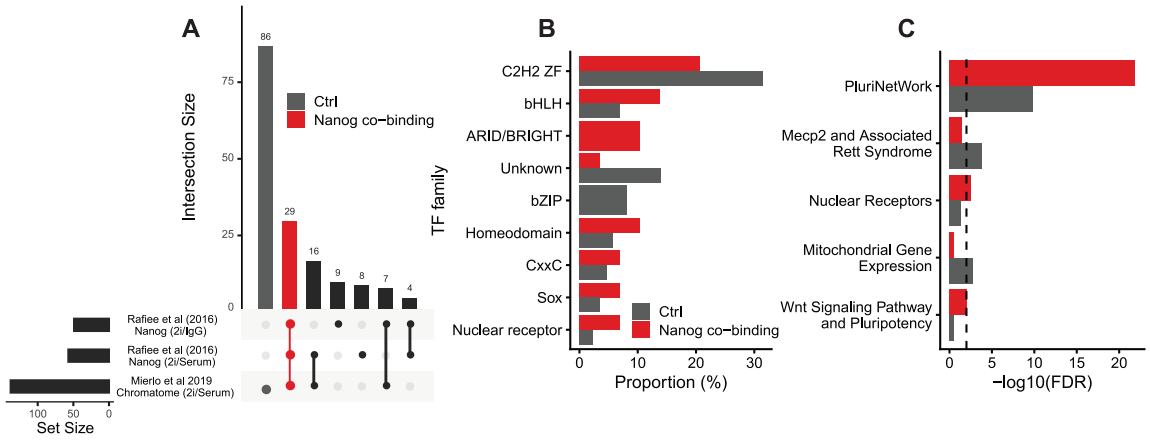


Fig. 2.19: Proteomics meta-analysis of published Nanog co-binding proteins and ESC Chromatome. (A) Upset plot showing overlap between TFs sets from three different proteomics datasets including chromatin-associated proteome (2i vs. serum) and NANOG co-binding proteins in either IP vs. no antibody or 2i vs. serum comparison [204, 205]. TFs were divided into two different classes, consensus NANOG co-binding TFs (red) and control TFs (grey). All other TFs were not considered in the downstream analysis. The top bar plot shows the number of TFs belonging to each set. The coloured dots underneath denote which set is being displayed. To the left of this, the total number of TFs taken from each study is shown in a bar plot. **(B)** Composition of TF families in in either set of NANOG co-binding (red) or control (grey) TFs). **(C)** Enrichment of WikiPathway terms in either set of NANOG co-binding (red) or control (grey) TFs. FDR from one-sided Fisher's exact test followed by Benjamini-Hochberg correction.

Having established a credible set of potential regulators of enhancer activity, we next aimed to associate each TF with a corresponding PWM. To achieve this, we made use of a recently published TF motif database comprised of 5,594 PWMs for 1,210 TFs [63]. However, the motif library contained multiple PWM matches for individual TFs; for example SOX2 was associated with 25 PWMs. Instead of using all TF-PWM links which would dramatically increase the number of features and would lead to a lot of redundancy, we developed a method to prune TF-PWM links in order to reduce the number of features. Briefly, each PWM was matched to NANOG binding site sequences and resulting scores were used to distinguish between active enhancers and inactive TFBS using the area under ROC curve (auROC) as measure of predictive power (for more details see Methods). We observed that PWMs belonging to the nuclear receptor family were highly predictive for active enhancers (Fig. 2.20A). More specific ESRRB ($auROC = 0.66$) and NR5A2 ($auROC = 0.65$) were among the TFs associated with the most predictive PWMs and also experimentally determined as NANOG co-binding TFs. PWMs associated with SOX2 ($auROC = 0.6$) and OCT4/POU5F1 ($auROC = 0.59$) also provided information to distinguish active from inactive sites. On the other hand, we found that PWMs belonging to the zinc finger and E2F

TF families showed higher association with inactive binding sites. Among them we found the following Nanog co-binding TFs: DNMT1 ($auROC = 0.61$), NR0B1 ($auROC = 0.59$), PATZ1 ($auROC = 0.59$), HES1 ($auROC = 0.59$) and TET1 ($auROC = 0.58$). In order to prune putative TF-PWM links, we selected, for each TF in our control and NANOG co-binding set, the PWM which maximises auROC and therefore showed the highest predictive potential (Fig. 2.20B). Additionally, we selected all PWMs $auROC > 0.6$ as a third feature set ("All pred") to later train another GBM model as control for this step. To summarise, our proteomics meta-analysis combined with PWM scoring and TF-PWM link pruning resulted in three distinct feature sets: all predictive PWMs ("All pred"), PWMs of NANOG co-binding TFs ("NANOG co-binding") and PWMs of chromatin associated TFs ("Ctrl"). To complete these feature sets, we added dinucleotides, in order to account for potential missing sequence features and make these models more comparable to the gkm-SVM and CNN approaches. Finally, we trained GBMs for each one of our three feature sets while also optimising model hyperparameters including the number of boosting iterations and tree depth (Supp. Fig.A.1A-C; see Methods).

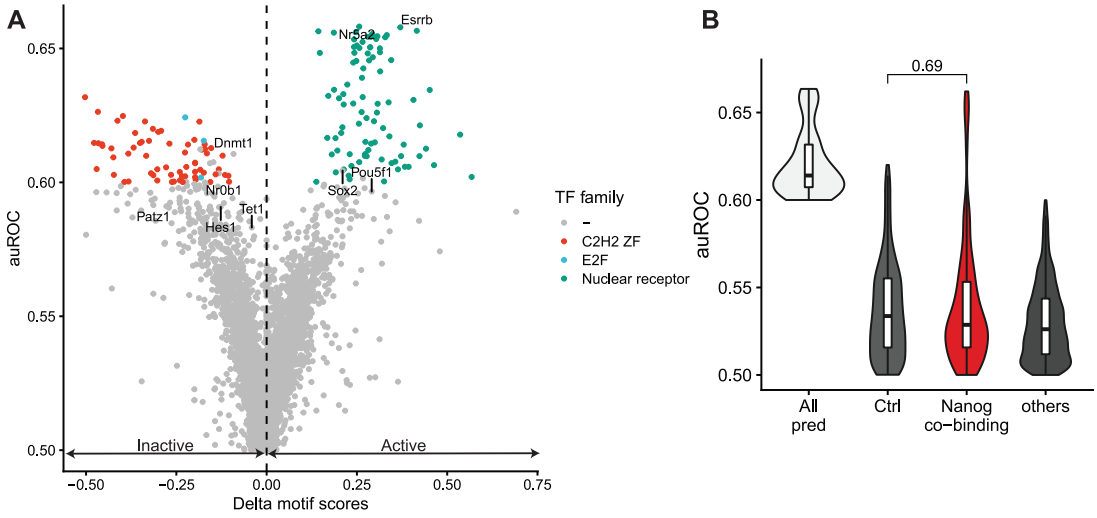


Fig. 2.20: Predictive power of single PWMs for enhancer activity of NANOG binding sites. (A) Volcano like plot illustrating the difference in median motif scores for each PWM against the predictive power, as measured by auROC. The best PWMs for NANOG co-binding TFs with $auROC > 0.58$ were labelled in the plot. Colours indicate the TF family of each predictive PWM ($auROC > 0.6$). **(B)** Comparison of predictive power as measured by auROC for our final three feature sets including: all predictive PWMs (light grey), control co-binding TFs (dark grey) and NANOG co-binding TFs (red). PWMs not selected as features were summarised as "others" (black). P-value from one-sided Wilcoxon test.

gkm-SVM hyperparameter selection and training

Second, we trained a gkm-SVM classifier to distinguish active enhancers from inactive binding sites. This method first computes a feature set based on gapped k-mers and then trains an SVM to distinguish classes [208]. Input features are obtained using three adjustable hyperparameters which can be optimised: k-mer length (L), gap length (K), and number of informative columns (NM; see Methods). We performed an exhaustive grid search to optimise these hyperparameters by 10-fold cross-validation training (Supp. Fig.A.2). The best parameters were selected based on maximisation of area under precision-recall curve (auPRC), which was $auPRC = 0.836$ for following parameters $L = 11$, $K = 6$, $NM = 3$. The resulting hyperparameters were used to train our final gkm-SVM model.

CNN training and hyperparameter selection

Third, we trained convolutional neural networks (CNNs) to classify active enhancers and inactive TFBS. One advantage of these models is that they are so-called end-to-end models meaning that no *a priori* feature selection was necessary, rather allowing direct learning from the DNA sequence itself [210]. We iteratively trained CNNs directly on the one-hot encoded Nanog peak sequence while simultaneously performing a hyperparameter search using Tree-structured Parzen Estimator (TPE) to find the best CNN architecture (see Methods) [209]. To avoid overfitting, we kept track of the loss and MCC for both training and validation datasets. Both datasets the loss and MCC plateaued on the validation set after approximately 80 epochs (Supp. Fig.A.3A). One draw-back with this training strategy is that the network weights were initialised randomly, thus necessitating a large training dataset to learn distinguishing features. Since we only had a limited number of examples, we made use of another training strategy called transfer learning which should reduce the amount of data required to train a CNN [210]. Briefly, a CNN was trained to distinguish Nanog binding peaks against dinucleotide shuffled background sequences (see Methods). Weights from the first convolutional layer of this successfully pre-trained CNN (test set: $auROC = 0.99$) were used to initialise weights of a new CNN which was then trained to distinguish active from inactive binding sites. The loss and MCC plateaued on the validation set during training of the best CNN architecture after approximately 40 epochs (Supp. Fig.A.3B).

GBM, gkm-SVM and CNN evaluation on held-out test set

We evaluated the above trained models on the same held-out test set and compared them using five different metrics including area under precision recall curve (auPRC), Matthews

correlation coefficient (MCC), balanced accuracy, sensitivity and specificity (see Methods) [194]. Comparison of all models using the auPRC as metric showed that both CNN models outperformed all others in terms of successfully classifying active enhancers (Fig. 2.21A). The NANOG co-binding TF GBM outperformed the gkm-SVM and the other GBMs in this comparison. However, this metric is focused on successfully recovering active enhancers (true positives) while not considering the number of successfully classified inactive binding sites (true negatives). Therefore, we computed the MCC for each model since this metric takes the entire confusion matrix into account (see Methods). This showed that the transfer-learning CNN performed best in accurately distinguishing active enhancers from inactive TFBS (Fig. 2.21B). In contrast, the GBM trained on the control TF set performed worst. The standard CNN and gkm-SVM models performed fairly well ($MCC > 0.4$). Two caveats of MCC as a metric are that it might strongly depend on the applied probability or score for gkm-SVM thresholds and that it might be misleading if the label distribution dataset is too imbalanced ($y_0 = 730$, $y_1 = 852$). Therefore, we additionally computed the balanced accuracy as another performance measure [194]. The gkm-SVM showed the highest and the transfer-learning CNN model the second highest balanced accuracy (Fig. 2.21C). Again, the GBM trained on the control TFs showed lowest accuracy. To better understand MCC and balanced accuracy metrics, we also computed the sensitivity and specificity for all models. All three GBM models showed the highest sensitivities as well as the lowest specificities, meaning that these models were better in identifying true active enhancers than inactive binding sites (Fig. 2.21C). On the other hand, the gkm-SVM model showed higher specificity than sensitivity, meaning that it might have learned to accurately identify inactive binding sites. We found that our CNN models showed the most balanced behaviour between sensitivity and specificity, providing further support for these models. It is important to keep in mind that balanced accuracy, sensitivity and specificity also depend on chosen probability or score threshold. Taken together, our transfer learning CNN model outperformed all other models in terms of auPRC as well as MCC and showed the best trade-off between sensitivity and specificity. Nevertheless, all trained models showed that they have some predictive power and that sequence features provide enough information to distinguish active enhancers from inactive TFBs.

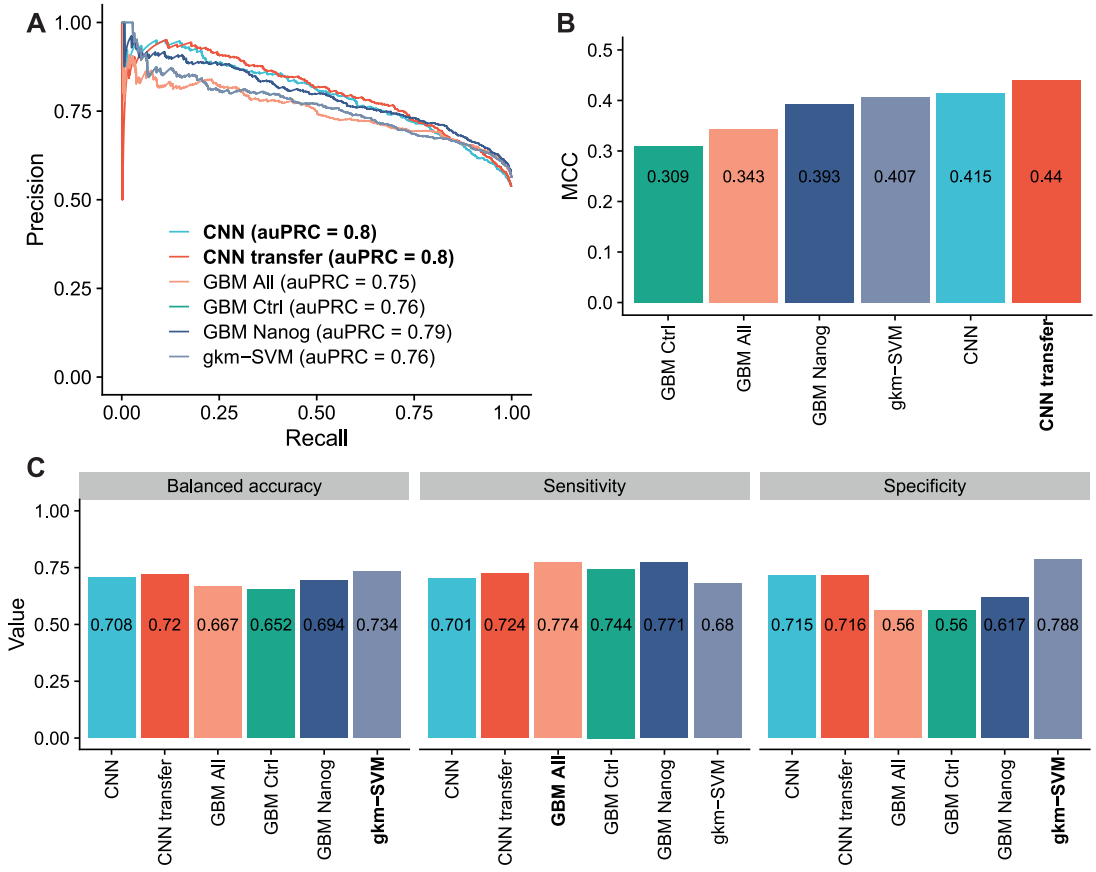


Fig. 2.21: Performance benchmark on hold-out test set of classification models using different metrics. **(A)** Precision-recall curves for all trained models including CNN (light blue), transfer-learning CNN (red), GBM all predictive features (orange), GBM control co-binding TFs (green), GBM NANOG co-binding TFs (dark blue), gkm-SVM (violet). Precision-recall curves were computed on held-out test-set. Area under precision recall curve (auPRC) for each model is shown in the legend. **(B)** Matthews correlation coefficient (MCC) for all of trained models evaluated on held-out test set. MCC labels are shown rounded to three digits after decimal point. **(C)** Balanced accuracy, sensitivity (true positive rate) and specificity (true negative rate) for all trained models evaluated on held-out test set. Values for performance metrics labels are shown rounded to three digits after decimal point.

GBM feature importance of NANOG co-binding TFs

Due to the high interpretability of GBMs, we made use of our GBM Nanog co-binding TF model to better understand which features/TFs were the most important predictors of enhancer activity [211]. To this end, we computed an importance score for each feature based on its relative influence on the final prediction (see Methods). This revealed that the ESRRB motif was by far the most important feature for the GBM model (Fig. 2.22A). We also found the TBX3, SOX2, PATZ1, OCT4 and NANOG motifs. In contrast, only the two dinucleotides

CG and GG, were among the top 15 most important features. To better understand the connection between the top three features and enhancer activity, we plotted motif scores for each against quintiles of the absolute STARR-seq signal. In case of ESRRB, we found that active enhancers showed generally contained better motif matches than inactive binding sites and that motif scores were increased with enhancer activity (Fig. 2.22B). The SOX2 motif exhibited comparably high matching scores in all active enhancer quintiles which were all higher than those of inactive binding sites. These observations are in line with our comparison of SOX2 ChIP-seq and CUT&RUN for active/inactive sites, showing a higher signal for active enhancers but no correlation with STARR-seq signal (Fig. 2.8, Fig. 2.9). In contrast, the TBX3 motif was better at matching sequences from inactive binding sites but did not show an observable connection between motif scores and STARR-seq signal.

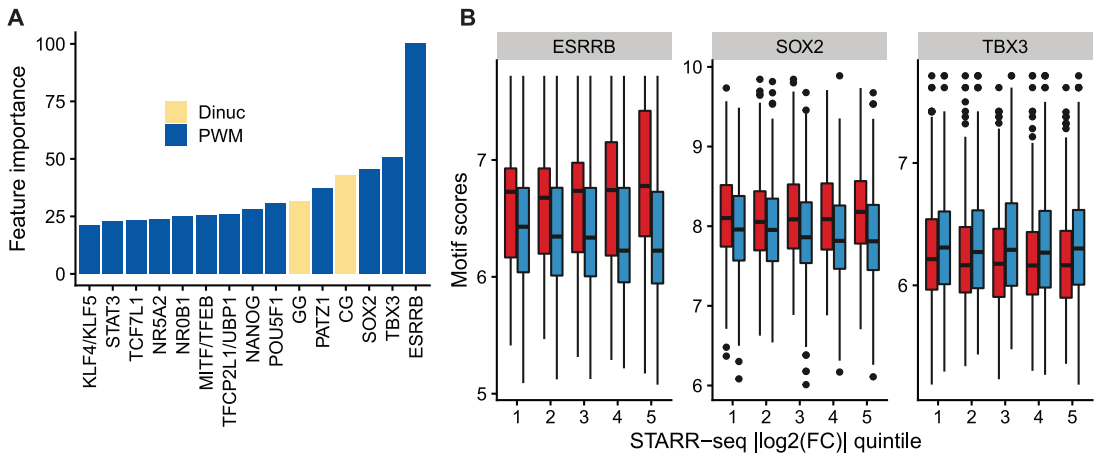


Fig. 2.22: Feature importance ranking for GBM trained on NANOG co-binding TFs. (A) Feature importance ranking showing the 15 most important features. Colour indicates whether the feature was a PWM (blue) or dinucleotide frequency (yellow). **(B)** Motif score distributions for the top three most important PWM features including ESRRB, SOX2 and TBX3. Motif scores are shown for each STARR-seq signal quintile ($|log2(FC)|$).

Having observed comparable trends between the SOX2 motif and sequencing data, we asked if this would held true for ESRRB as well as TBX3. Using ChIP-seq data we could confirm greater ESRRB binding at active enhancers than inactive TFBS could confirm higher binding of ESRRB at active enhancers following the observed motif trend ($median_{active} = 2.4$, $median_{inactive} = 1.49$, $p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.23B). In contrast, TBX3 ChIP-seq revealed significant higher average binding to inactive TFBS ($median_{active} = 0.72$, $median_{inactive} = 0.86$, $p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.23D).
In summary, understanding which features are important for the trained GBM to distinguish active enhancers and inactive TFBS revealed ESRRB, SOX2 and TBX3 motifs as major

contributors. This finding was further supported by their binding profiles measured by ChIP-seq.

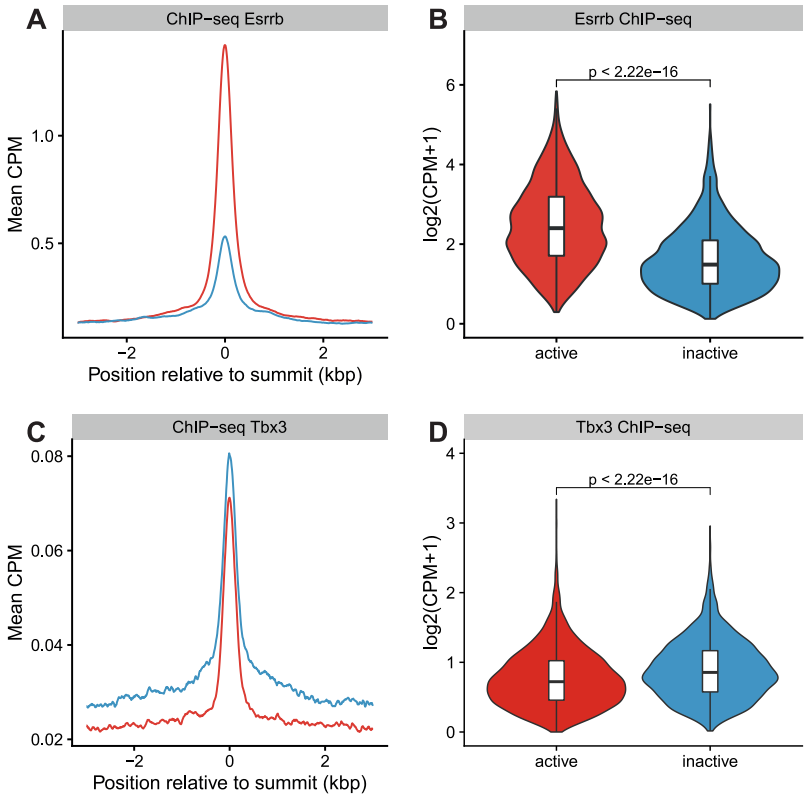


Fig. 2.23: ESRRB and TBX3 binding at active enhancers and inactive TFBs. Metaprofile showing average binding signal for **(A)** ESRRB and**(C)** TBX3 ChIP-seq. Violin plot showing comparison of TF enrichment within active enhancers (red) and inactive binding sites (blue) for **(B)** ESRRB and **(D)** TBX3 ChIP-seq. ChIP-seq signals were normalised to counts per million (CPM). P-values from one-sided Wilcoxon test.

Interpretation of CNN provides base-pair level importance score

We next sought to identify the sections within the entire 500bp sequence of active enhancers and inactive TFBS that are important for mediating enhancer activity. We used our CNN model for this, since it learned from the one-hot encoded raw sequences (see Methods). This allowed us to apply a back-propagation approach, as implemented by DeepLIFT, to assign an importance score each base-pair: a positive score if it was predictive for enhancer activity and negative if it was predictive for inactive TFBS [212].

First, we investigated if computed DeepLIFT scores were in line with our previous observations that ESRRB is an important predictor of enhancer activity (Fig. 2.22). Therefore,

we made use of recently published ESRRB and STAT3 ChIP-exo data (available from the same study) which provides TF binding at motif level resolution [254]. Obtained ESRRB ChIP-exo peaks were significantly more enriched in active enhancers compared with inactive TFBS ($p < 2.2 \cdot 10^{-16}$, one-sided Fisher's exact test). Due to the high resolution of these peaks, we were able to compute average DeepLIFT scores for $\pm 50bp$ around ESRRB as well as STAT3 summits. This revealed central enrichment of importance scores around ESRRB ChIP-exo summits, independently of whether they overlapped with inactive TFBS or active enhancers (Fig. 2.24A). In contrast, STAT3 did not show any enrichment of DeepLIFT scores over background levels in neither active enhancer nor inactive TFBS. Further, active enhancers showed significantly higher DeepLIFT scores around ESRRB summits than inactive TFBS ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 2.24B). We found for that summed DeepLIFT scores for STAT3 were on average close to zero for both peak classes. Thus, this analysis confirmed the validity of our base-pair importance scores and further supported ESRRB as important predictor of enhancer activity.

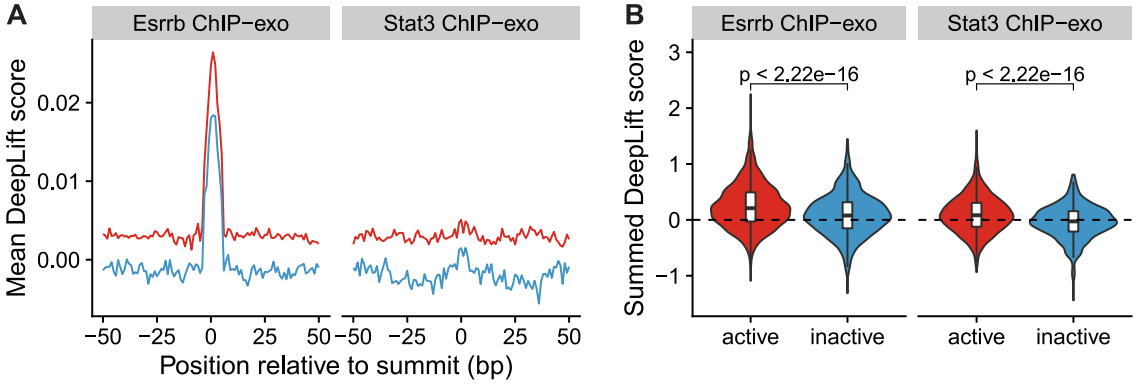


Fig. 2.24: DeepLIFT importance scores relative to ChIP-exo summits. **(A)** Average DeepLIFT importance scores relative to ESRRB (left) and STAT3 (right) ChIP-exo summits. ChIP-exo summits were stratified according to their overlap with active enhancers (red) or inactive TFBS (blue). **(B)** Violin plot showing summed DeepLIFT importance scores for $\pm 15bp$ around ChIP-exo summits of ESRRB (left) and STAT3 (right). P-value from one-sided Wilcoxon test.

To systematically analyse base-pair level importance score further, we performed motif discovery with TF-MoDisco [213]. Briefly, TF-MoDisco uses DeepLIFT scores to define important sub-sequences within each peak, followed by clustering of these sequences and compilation of similar ones into motifs (see Methods). This revealed six TF motifs associated with inactive TFBS and eight for active enhancers (Fig. 2.25). We clustered these de novo motifs revealing six distinct clusters including three for active enhancers and three for inactive

TFBS. Subsequent motif annotation uncovered that our three active enhancer motif clusters were comprised of nuclear receptor TFs (e.g. ESRRB, NR5A2), TEAD TFs and repeat motifs (Fig. 2.25). Inactive TFBS motif clusters included T-box TFs (e.g. TBX3), Homeobox/ARID TFs (e.g. NANOG, ARID3A) and zinc-finger/poly(G)-repeat motifs. Interestingly, we found four different nuclear receptor motifs for ESRRB, which is in line with our previous observation that ESRRB was the most important factor for predicting active enhancers. This analysis also recovered a motif associated with TBX3 which we previously uncovered as the second most important TF and being associated with inactive TFBS.

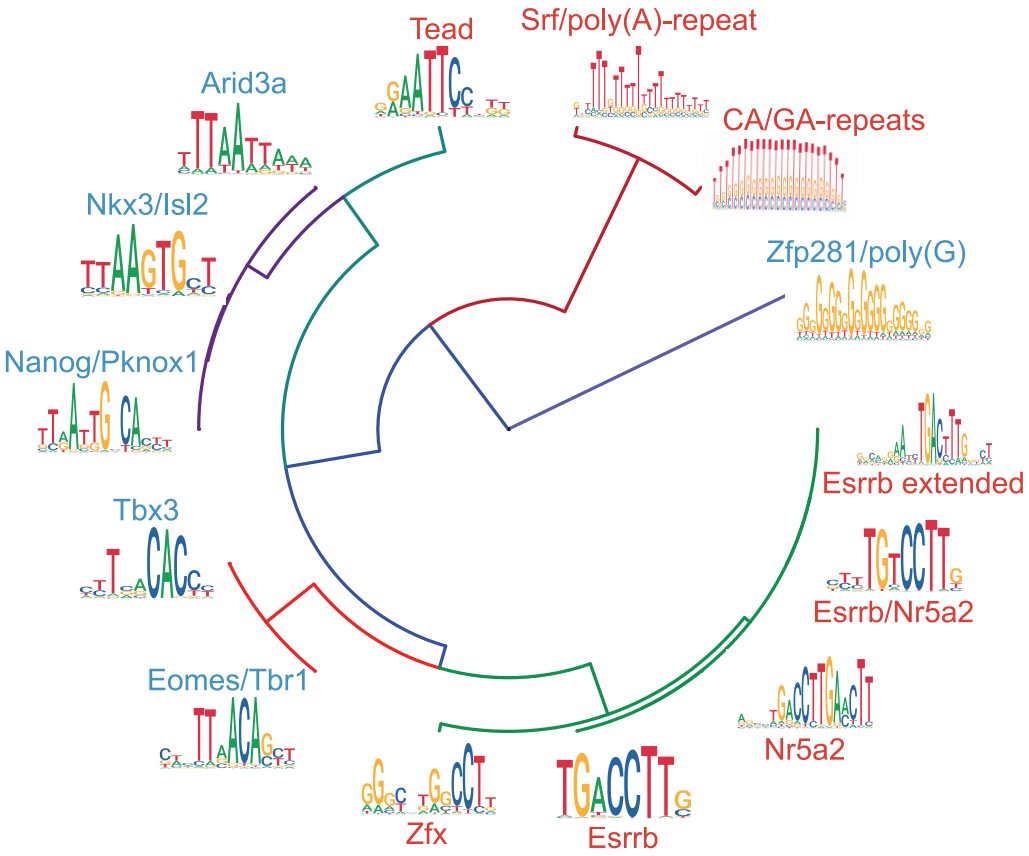


Fig. 2.25: TF-MoDisco de novo discovered TF motifs from DeepLIFT importance scores. TF-MoDisco de novo discovered motifs based on DeepLIFT importance scores were clustered, annotated and visualised as a circular dendrogram. Discovered motifs could be assigned to six different clusters including nuclear receptors (green), T-box (red), Homeobox/ARID (purple), TEAD (cyan), CA/GA/poly(A) repeats and zinc finger/poly(G) repeats. The names of TFs next to their associated motif are coloured according to their predictive power for active enhancers (red) or inactive TFBS (blue).

Next, we asked which of the TF-MoDisco de novo motifs might be the most important for distinguishing between active and inactive sites. To this end, we applied a similar strategy

as for in the GBM feature selection procedure and used each PWM as single classifier to compute auROCs as measure of predictive power (Fig. 2.26). This analysis revealed that the ZFP281/poly(G)-repeat and TBX3 were the most important features to predict inactive TFBS, in line with our previous observations that TBX3 was the second most important GBM feature. On the other hand, all four discovered nuclear receptor motifs were amongst the most important features, followed by the poly(A) and CA/GA-repeat motifs (Fig. 2.26). However, the nuclear receptor motif resembling most the canonical ESRRB motif showed the highest predictive power amongst active enhancer features. This was again in line with our previous observations showing ESRRB as most important GBM feature (Fig. 2.22)

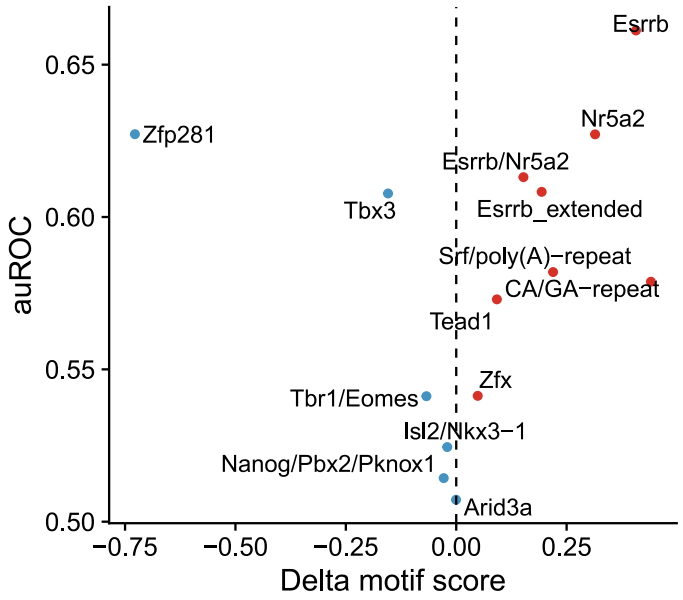


Fig. 2.26: Feature importance of TF-MoDisco de novo discovered TF motifs. Volcano like plot visualising the difference in median PWM scores against predictive power as measured by auROC for each discovered TF-MoDisco motif. Points were coloured according to their association with active enhancers (red) or inactive TFBS (blue).

Having observed four different de novo motifs discovered using TF-MoDisco, we asked which of them would be engaged by ESRRB in the cell. To answer this we made use again of the ESRRB ChIP-exo data introduced earlier by plotting the signal relative to TF-MoDisco discovered motifs. We performed this for all four nuclear receptor motifs and five other de novo motifs as control. This revealed the typical bi-modal ChIP-exo enrichment at all four nuclear receptor motifs confirming ESRRB binding in ESCs (Fig. 2.27). In contrast, none of the controls revealed this typical pattern, and even three out of five motifs (Arid3a, CA/GA-repeats, poly(A)-repeats) showed clear depletion of ChIP-exo signal. Thus,

providing evidence for the importance of these nuclear receptor motifs within the normal chromosomal context of these active pluripotency enhancers.

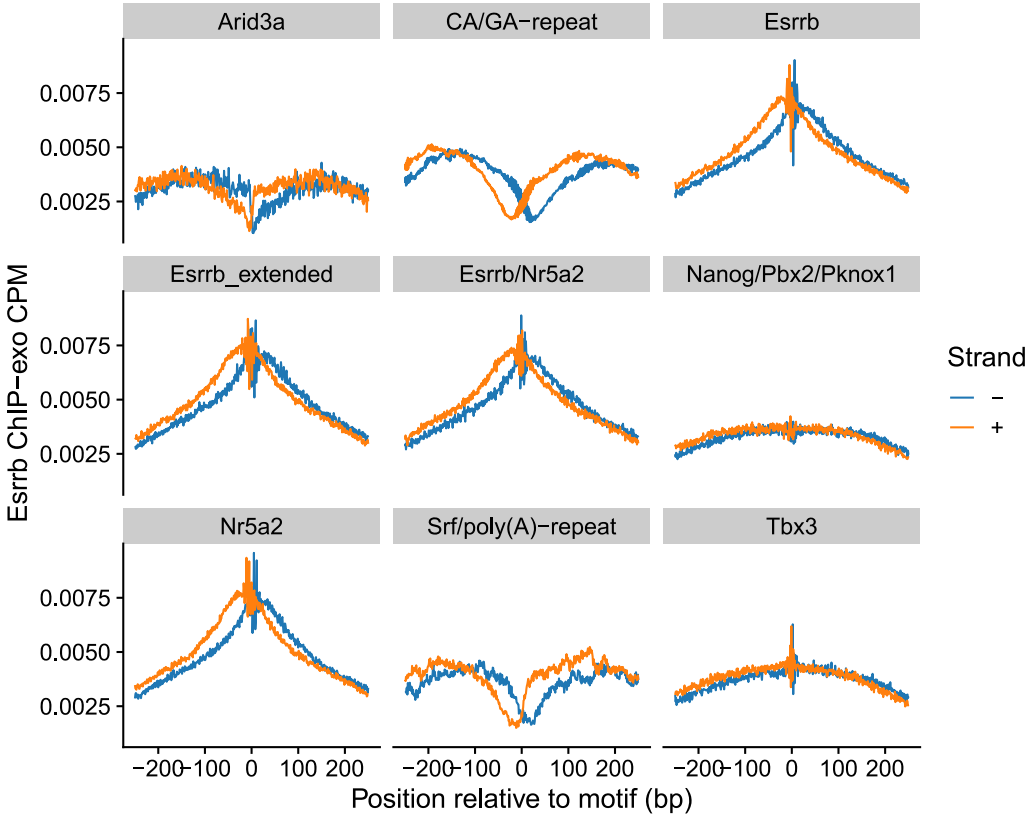


Fig. 2.27: Binding of ESRRB to TF-MoDisco de novo discovered motifs. Average ESRRB ChIP-exo signal visualised as metaprofile for $\pm 250bp$ around respective motif matches. ChIP-exo signal was split into forward (orange) and reverse strand (blue).

Taken together, this analysis provides us with a meaningful base-pair resolution importance score. In addition, it revealed previously known and unknown TF motifs involved in enhancer activity as well as NANOG TFBS inactivity. Ultimately, we found additional support for the nuclear receptor ESRRB being the most important feature for predicting enhancer activity.

2.3.11 Knockout of *Esrrb* leads to loss of H3K27ac at active Nanog binding enhancers

The predictive modelling to distinguish active enhancers from inactive TFBS highlighted the ESRRB binding motif as being the most important sequence feature for *bona fide* enhancer elements. The ESRRB motif was the most predictive feature in our feature selection procedure (Fig. 2.20), most important variable for the GBM model (Fig. 2.22), our CNN

model learned multiple ESRRB motifs which all were predictive for enhancer activity (Fig. 2.25, 2.26), TF ChIP-seq peak enrichment showed ESRRB as one of the top factors (Fig. 2.11) and ESRRB ChIP-seq analysis confirmed differential engagement at active enhancers compared with inactive TFBS (Fig. 2.23). However, the nuclear factor family shares a high degree of motif similarity especially between NR5A2 and ESRRB [255]. To establish a causal connection between enhancer activity and ESRRB binding we re-analysed publicly available H3K27ac ChIP-seq from *Esrrb* and *Nr5a2* inducible knockout (KO) ESCs [199].

This confirmed the enrichment of H3K27ac on active enhancers compared with inactive TFBS for all profiled conditions (Fig. 2.28A). Induced KO of *Nr5a2* showed only a marginal effect on average H3K27ac levels and only affected a small number of peaks ($median_{active} = 0.51$, $median_{inactive} = 0.53$; Fig. 2.28A, B). In contrast, *Esrrb* KO was accompanied by a reduction in average H3K27ac enrichment at active enhancers to levels observed at inactive binding sites (Fig. 2.28A). Loss of H3K27ac was observable at the majority of active enhancers and these loci lost more in H3K27ac enrichment upon KO compared with inactive TFBS (ESRRB 2i day3: $median_{active} = 0.33$, $median_{inactive} = 0$; ESRRB 2i LT: $median_{active} = 0.57$, $median_{inactive} = 0.11$; $p < 2.2 \cdot 10^{-16}$, one-sided Wilcoxon test). This effect was stronger in the ESC 2i long-term condition compared with only the 3 day culture, which might be explained by a more synchronised cell population. In summary, we present causal evidence for ESRRB being one of the most important factors for activity of pluripotency enhancers in ground-state (2i) ESCs. Our results imply that ESRRB is regulating enhancer function via co-activator recruitment resulting either in the maintenance or establishment of H3K27ac.

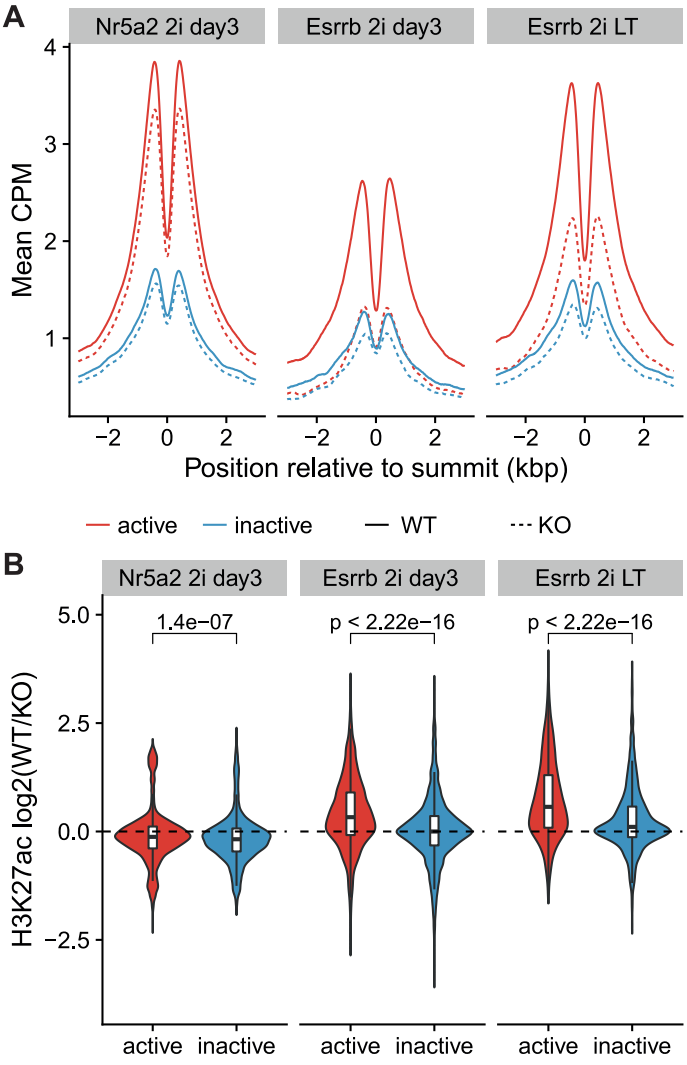


Fig. 2.28: H3K27ac ChIP-seq from ESC with inducible knockouts of either *Nr5a2* or *Esrrb*. **(A)** Metaprofiles showing H3K27ac ChIP-seq enrichment over active enhancer (red) or inactive TFBS (blue) summits from WT ESC (solid line) and *Nr5a2*/*Esrrb* induced KO (dotted line). ChIP-seq was performed on two different ESC conditions including ESC cultured for three days under 2i condition (2i day3) and 2i long-term culture (2i LT). **(B)** Violin plot showing quantification of H3K27ac changes between ESC WT and either induced *Nr5a2* or *Esrrb* KO within active enhancers (red) and inactive TFBS (blue). CPMs over ±250bp window for each peak were computed. P-values derived from one-sided Wilcoxon test.

2.4 Discussion

In this chapter, we examined how *bona fide* enhancer activity is encoded in Nanog binding sites and defined the key DNA sequence determinants of this. To this end, we made use of a recently published method combining ChIP with the high-throughput reporter assay STARR-seq [178]. This uncovered two classes of NANOG TFBS, one corresponding to *bona fide* transcriptional enhancers and a second class of inactive TFBS. We comprehensively characterised these two classes of NANOG binding sites by TF binding, chromatin landscape, chromatin-chromatin interaction, target gene expression and DNA sequence composition.

In an initial step, we applied a stringent procedure to define a high-confidence consensus peak set of NANOG binding sites. These peaks exhibited high levels of NANOG ChIP-seq enrichment independent of study or condition as well as ChIP-independent binding signal from NANOG CUT&RUN. Additionally, we observed higher enrichment of NANOG and NANOG-OCT4-SOX2 composite motifs than in the filtered peaks. However, these peaks might still be composed of multiple NANOG binding sites. One improvement to our experimental strategy could be made by fine-mapping NANOG binding sites to motif level using high-resolution data such as ChIP-exo or ChIP-nexus [117, 118].
An important step to identify high-confidence consensus peaks is the determination of a coverage threshold to filter out background regions. Therefore, we sampled a non-overlapping set of random regions by randomly selecting them from the remaining mouse genome. However, this analysis step could be improved by matching important co-variates such as GC-content in the randomly drawn peak set. Additionally, this could be constrained to open-chromatin regions, for e.g. from ATAC-seq, not overlapping the binding sites of interest.
Having stringently defined NANOG binding sites, we performed a differential analysis between the vector library and STARR-seq from ground-state (2i) ESC. This revealed 3,510 *bona fide* active enhancers and 2,838 NANOG bound TFBS which did not show enhancer activity which we termed inactive TFBS (Fig. 2.6A). Additional support for our observations came from a recent study profiling enhancer activity using NANOG/OCT4/H3K27ac/H3K4me1 ChIP STARR-seq in human ESC also revealing an inactive fraction of Nanog binding sites [180]. However, the largest fraction of our Nanog consensus peaks (n=7,299) did not show significant differences before and after transfection. Multiple experimental parameters might explain this. First, we used a version of the mammalian STARR-seq vector containing SCP1 as minimal promoter. A recent publication showed that this construct design results in

competing transcription initiation events from the minimal promoter and origin of replication (ORI) leading to a reduction in STARR-seq signal [176]. Second, the largest proportion of signal in STARR-seq experiments originates from fragment level enhancer self-transcription resulting in many duplicates of the same fragment. However, we were unable to distinguish these from PCR duplicates and had to use unique fragments. Introduction of unique molecular identifiers (UMIs) would enable discrimination between expressed fragments and from PCR duplicates, and therefore greatly increase the signal yield [219, 184]. Moreover, introduction of UMIs would allow us to consider each cloned fragment as separate measurement and therefore would give us higher resolution by tiling the enhancer regions [256, 179]. Ultimately, additional biological replicates would increase the statistical power to discriminate weakly transcribed enhancers from inactive TFBS.

Nevertheless, the active enhancers we identified included previously described enhancers involved in regulating important pluripotency genes such as *Nanog*, *Sox2* and *Esrrb* [223, 257, 60]. Interestingly, we found that our inactive NANOG TFBS overlapped with previously described developmental enhancers. For example, a NANOG TFBS distal to *Hoxa1* which becomes active upon RA treatment and was previously described as being involved in neuro-ectoderm development [222]. Moreover, inactive Nanog TFBS were located close to pluripotency and developmental TFs, for example *Foxd3*, *Gata6* and *Foxa2* [221, 258]. Further support for this comes from a recent publication which suggested that enhancer elements in more differentiated cell types were pre-marked by pluripotency factors including NANOG [229].

We characterised active enhancers and inactive TFBS in terms of TF-binding, chromatin landscape, chromatin-chromatin interaction and target-gene expression. For active enhancers, we recovered all previously described enhancer hallmarks including high TF-engagement, DNA-accessibility and nucleosome-depletion, enrichment of H3K4me1 as well as H3K27ac, bi-directional eRNA expression and enhanced transcription of target genes [161, 42, 141]. In contrast, inactive TFBS were characterised by lower TF-engagement, higher nucleosome-density, enrichment of H3K4me1 and the repressive Polycomb mark H3K27me3, little nascent eRNA transcription and weak expression of target genes. In line with our observations, a recent publication provided evidence that NANOG not only occupies pluripotency enhancers but also sustains repressive H3K27me3 at enhancers of important developmental regulators [259].

Having observed such substantial differences between the two classes of NANOG binding sites, we hypothesised that their states must be encoded in their underlying DNA sequences. To decode the regulatory logic, we established a machine learning pipeline based on three different algorithms including GBM, gkm-SVM and CNN [211, 208, 210]. This included a novel approach for selecting biologically relevant sequence features based on proteomics as well as TF motif data. Training of these models to discriminate active enhancers from inactive binding sites confirmed the predictive power of sequence-encoded information regardless of the algorithm employed. We took advantage of trained GBM and CNN models to gain a better understanding of the important features. Initial analysis of the selected GBM features highlighted ESRRB as the most important for active enhancers and TBX3 as the most important one for inactive TFBS (Fig. 2.22). Differential binding of these factors was confirmed by analysis of publicly available ChIP-seq data (Fig. 2.23). Since we trained our CNN end-to-end, using one-hot encoded DNA sequence as input, we could derive a base-pair level importance score using DeepLIFT [212]. Clustering of these importance scores for de novo motif discovery revealed known and unknown motifs associated with active enhancers and inactive binding sites. The most prominent motifs were associated with the nuclear receptor family which were annotated as ESRRB (Fig. 2.25). Amongst these de novo discovered motifs, we also found one motif associated with TBX3. Subsequently computed feature importance scores based on these de novo discovered motifs confirmed ESRRB and TBX3 as important variables, but also revealed ZFP281/poly(G)-repeat as an important feature of inactive TFBS (Fig. 2.26).

All our analyses pointed towards the nuclear receptor ESRRB as being the single most important predictor of enhancer activity. This is supported in the published literature highlighting the role of ESRRB in transcriptional regulation and its importance to induce or maintain the pluripotent ESC cell phenotype. Evidence for the important role of ESRRB to maintain the naive ESC state was provided by a recent publication showing that downregulation of ESRRB was associated with loss of naive and induction of primed pluripotency [260]. Moreover, it was shown that ESRRB promotes reprogramming of somatic cells to induced pluripotent stem cells (iPSCs) by pioneering silenced enhancer elements [251, 261]. In terms of transcriptional regulation, a study performing systematic shRNA knockdown (KD) of 100 TFs in ESCs revealed *Esrrb* KD as causing the most severe effects with $2,500$ differentially expressed genes [262]. These marked changes could be explained by genome-wide lose of pluripotency enhancer activity resulting in differentiation. Additional evidence for this was provided by two recently published papers which showed that knockdown or knockout

of *Esrrb* is accompanied by loss of the mediator co-factor and H3K27ac from enhancer elements [263, 199]. Another study showed that ESRRB bound enhancers associated with important pluripotency genes (*Esrrb*, *Klf4*, *Nanog*, *Sox2*) lose their enhancer activity upon *Esrrb* KD, as measured by luciferase reporter assay [264]. In case of *Klf4* associated enhancers, another study even showed that loss of ESRRB motifs resulted in down-regulation of *Klf4* mRNA levels [254]. In summary, there is substantial experimentally derived evidence in the published literature supporting our results of ESRRB being the most important TF for pluripotency enhancer activity.

Our analysis also revealed TBX3 and ZFP281 as important features for predicting inactive TFBs. Interestingly, KD of these TFs in ESC resulted in very few changes (*Zfp281* KD 11 genes, *Tbx3* KD 34 genes) [262]. TBX3 was initially described as one of the TFs involved in leukaemia inhibitory factor (LIF) signalling mediating pluripotency [265]. However, a more recent publication described TBX3 as dispensable for the induction and maintenance of pluripotency [266]. Several studies reported TBX3 as being involved in pluripotency exit control and directing differentiation towards either a mesodermal or endodermal phenotype [267, 268, 269]. It was suggested that TBX3 might achieve this by recruitment of the histone demethylase JMJD3 to remove repressive H3K27me3 from developmental enhancers [268]. This is in line with our observations of inactive TFBS being enriched for H3K27me3 and close to developmental regulators such as *Eomes* (Fig. 2.6, 2.12). ZFP281 has been described in the literature as an activator as well as repressor that is involved in exit from pluripotency as well as regulating primed pluripotency [270, 271]. As potential mechanisms, it was suggested that ZFP281 might be able to recruit histone methylases and de-/acetylases to its target sites [272, 270, 271]. One example described in a recent paper identifies an enhancer in the *Nodal* locus that is bound by ZFP281 which overlaps our inactive TFBS (data not shown) [271]. These findings support our hypothesis that inactive NANOG TFBS become active enhancers at some point during development and are already bookmarked by NANOG in ground-state ESC.

In summary, in this chapter we comprehensively described and characterised two classes of NANOG binding sites including *bona fide* enhancer elements and inactive TFBS. Predictive modelling of these two classes was able to uncover distinct regulators such as ESRRB for active enhancers and TBX3/ZFP281 for inactive TFBS. However, additional wet-lab experiments are necessary to validate our predictions. In the future, application of ChIP STARR-seq in combination with machine learning approaches might help uncover more

TFs with dual functions, characterise key determinants of enhancer activity and uncover important regulators of certain phenotypes.

3

Assessing condition-specific enhancer activity of NANOG binding sites

Mouse embryonic stem cells can be cultured using either primed (serum) or naive (2i) conditions resulting in two distinct phenotypes. The 2i condition is considered closer to the "ground state" of pluripotency. By contrast, ESC cultured in serum show a phenotype primed for differentiation into other cell types. Nevertheless, ESCs in both conditions remain pluripotent and rely for this on the gene regulatory circuit involving NANOG, OCT4 and SOX2. However, understanding of how differential gene expression is encoded in the genome by binding sites of these factors remains elusive. Here, we examine how differential enhancer activities between the primed and naive conditions are encoded in the DNA sequence underlying Nanog binding sites. To this end, we combined ChIP with the high-throughput enhancer assay STARR-seq to test for enhancer activities. Comparison of Nanog ChIP STARR-seq data between 2i and serum conditions uncovered condition specific enhancer elements. We characterised these elements using publicly available data recovering previously described enhancer hallmarks such as enrichment of H3K27ac in a condition specific manner. Moreover, we found that changes in chromatin accessibilities between 2i and serum conditions are the major predictors of differential enhancer activity. Finally, we applied predictive modelling methods, including gradient boosted trees and support vector machines, to gain a better understanding of how the observed differential enhancer activity is encoded in the DNA sequence. This revealed TCF4, TCF7L1 and ESRRB as major predictors for 2i enhancer activity and ZIC3 for serum. Integration of publicly available ChIP-seq confirmed the differential binding of these factors at these enhancers. Taken together, these analyses demonstrate how the combination of enhancer reporter assays paired with machine learning can uncover sequence determinants and identify candidate TFs as potential regulators of condition specific enhancer activity.

Declaration

The work presented in this chapter was a collaborative work within the Luscombe Lab. Mahmoud-Reza Rafiee performed wet lab experiments including generation of ChIP STARR-seq and Nanog ChIP-seq data.

I (= Sebastian Steinhauser) analysed all high-throughput sequencing data including ChIP STARR-seq, ChIP-seq, GRO-seq, ATAC-seq and additional data sources presented in this chapter. All figures in this chapter were made by myself and plots for them were generated with code I implemented.

3.1 Introduction

Embryonic stem cells (ESCs) are pluripotent, meaning they are able to self-renew and differentiate into the full range of cell types characteristic of the embryo [225]. Hence these pluripotent cells closely resemble properties of cells found in embryos at early stages of development. These ESCs can be maintained in cell culture using diverse conditions, each inducing slightly different phenotypic properties. These culture conditions can be broadly distinguished into two major categories, depending on the media additives: two kinase inhibitors (2i) or serum [225]. Both are usually supplemented with Leukemia Inhibitory Factor (LIF) which induces the JAK-STAT3 signalling pathway resulting in the activation of self-renewal genes [273]. The 2i conditions refers to the use of kinase inhibitors PD0325901 (PD), which inhibits mitogen-activated protein kinase (MAPK) leading to impaired ERK signalling, and CHIR99021 (CHIR) inhibiting Glycogen synthase kinase 3 (GSK-3) resulting in induction of WNT signalling [274]. ESCs grown under 2i conditions more closely resemble cells from the preimplantation epiblast and therefore are referred to as ground-state/naive ESC. These naive ESCs display distinct molecular properties such as genome-wide hypomethylation and low levels of H3K27me3 at developmentally important genes [275, 276, 277]. In contrast, ESCs grown in the presence of serum show a phenotype primed for differentiation ('primed ESC'), based on reinforcing pluripotency via induction of STAT3 signalling but without blocking differentiation signals from FGF/ERK signalling pathways [273]. These more primed ESCs display higher cell-to-cell variability in cell culture and have poorer performance in standard pluripotency assays compared with 2i [273, 274]. They also exhibit genome-wide DNA methylation and higher levels of H3K27me3 at developmental genes [275, 276, 277].
Many gene expression studies comparing serum and 2i have been performed to provide mechanistic insights into how these phenotypic differences arise. They uncovered thousands of differentially expressed genes between the conditions (2i: 1500 and serum 2000 upregulated genes) [275, 199]. 2i expressed genes were associated with self-renewal and stem cell maintenance genes and serum expressed genes with developmental/differentiation functions. To find possible explanations for the observed difference in gene expression, studies investigated the various layers of transcriptional regulation including altered chromatin conformation and differential transcription factor (TF) binding. Examination of altered chromatin-chromatin interactions using capture Hi-C revealed only a few changes in long-range contacts and barely any differences at shorter ranges [199, 145]. In contrast, investigation of differential binding of core pluripotency factors (NANOG/SOX2/OCT4) displayed several thousand differentially bound sites between conditions [216]. However, relating these differential

binding sites to differentially expressed genes showed that the majority of sites were not associated with gene expression changes and only a weak correlation between binding and expression was found [216].

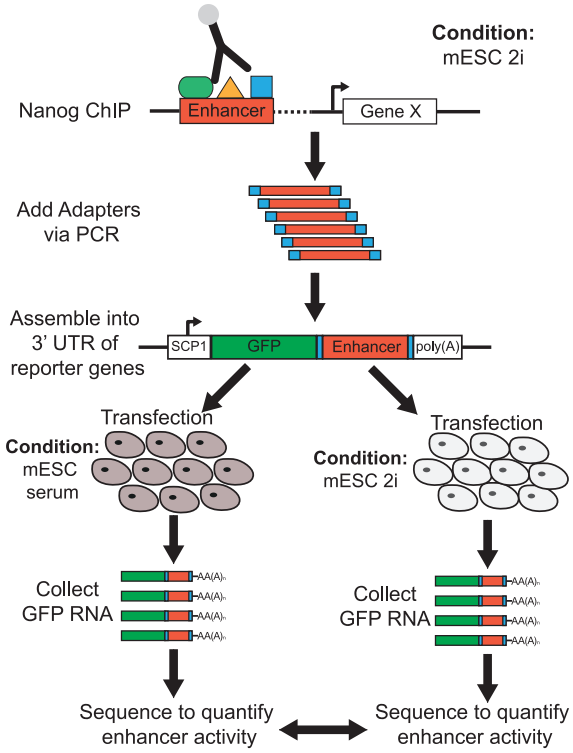


Fig. 3.1: ChIP STARR-seq experiment to determine condition-specific *bona fide* transcriptional enhancers from NANOG binding sites. First, a chromatin immunoprecipitation (ChIP) was performed to pull down NANOG cross-linked to DNA in naive (2i) mouse embryonic stem cells (mESCs). Subsequently, cross-linking was reversed, NANOG digested and pulled down DNA fragments were tagged with adapters via PCR. Tagged DNA fragments were cloned into the 3'UTR of GFP in the STARR-seq screening vector. In the screening vector, GFP is regulated by a SCP1 minimal promoter which is unable to induce GFP transcription without additional regulatory input. GFP will only be expressed if a *bona fide* enhancer element is cloned into the 3'UTR. The resulting STARR-seq vector library is transfected into either 2i or serum mESCs. Transcribed GFP mRNA is collected and prepared for sequencing. Therefore, active cloned enhancer sequences can be identified, as they will be in the 3'UTR of the expressed GFP mRNAs. To enable us to identify inactive NANOG-bound fragments, we also sequenced all the fragments cloned in the vector library. Cloned and transcribed fragments were target using the PCR added sequencing adapters [184]. We did not extract and sequence the transfected vector library. Finally, reads from control and STARR-seq libraries were aligned against *mm10* reference genome, quantified and differential analysis between 2i and serum STARR-seq was performed.

In this chapter, we hypothesised that the observed differential expression might be explained by transcriptional enhancers exhibiting differential activity between conditions. To this end,

we implemented a previously published assay combining chromatin immunoprecipitation (ChIP) with an episomal high-throughput enhancer assay called STARR-seq [178]. Briefly, NANOG ChIP was performed using 2i ESCs, resulting fragments were cloned into a STARR-seq vector library and transfected to serum and 2i grown ESCs (Fig. 3.1). It was hypothesised that exposure to the condition-specific transcriptional programs should lead to differential enhancer output in our enhancer assay. Subsequent, differential comparison of 2i and serum STARR-seq revealed condition-specific transcriptional enhancers. The resulting condition-specific enhancers were comprehensively characterised using publicly available datasets, leading to the recovery of many known enhancer hallmarks [161, 42]. For example, we could observe differential enrichment of the activating histone modification H3K27ac as well as its associated epigenetic writer enzyme P300 [83]. Finally, we addressed the question of how differential enhancer activities were encoded in the enhancers underlying DNA sequence using machine learning methods. This analysis uncovered TCF3/TCF7L1, TCF4/TCF7L2 and ESRRB motifs as predictors of 2i specific enhancer activity and ZIC3 as a predictor for serum specific enhancers. Together, in this chapter we identified condition-specific transcriptional enhancers, investigated causes of differential activity and revealed potential regulators.

3.2 Methods

3.2.1 Experimental procedures

Please see section 2.2.1. for brief descriptions of ChIP-seq and ChIP STARR-seq experiments.

3.2.2 Regulatory genomics sequencing data preprocessing and analysis

All sequencing datasets used in Chapter 3 were processed as described in Chapter 2. Preprocessing/Analysis sections and sample lists per sequencing technique for Chapter 3 are summarised in Table 3.1.

Table 3.1: Summary of sequencing methods, analysis methods and sample lists.

Sequencing methods	Analysis methods	Samples
ChIP-seq	Section 2.2.2	Supp. Table B.4
ChIP STARR-seq	Section 2.2.3	Supp. Table B.1
RNA-seq	Section 2.2.6	Supp. Table B.9
GRO-seq	Section 2.2.7	Supp. Table B.6
ATAC-seq	Section 2.2.8	Supp. Table B.6
Capture Hi-C	Section 2.2.10	Supp. Table B.7

Genome track plots shown in Figure 3.3 and 3.4 were plotted as described in Section 2.2.9.

3.2.3 Definition of 2i- and serum-specific enhancer elements

For this analysis we re-used the previously defined Nanog consensus peaks and pre-processed STARR-seq data sets, as described in Chapter 2 (see Section 2.2.2 and 2.2.3). Fragment abundance for all ChIP STARR-seq samples as well as the vector library was quantified using countOverlaps from the GenomicRanges library [191]. The resulting count matrix was used as input to edgeR to compare ESC 2i and serum conditions using the quasi-likelihood F-test procedure [195]. We filtered for differential peaks by applying an FDR threshold of 0.05. The remaining peaks were filtered for *bona fide* transcriptional enhancers by removing all peaks which displayed no enrichment over vector library in neither condition ($log2(STARR/vector) < 0$). Enhancers were then classified into 2i specific or serum specific according to the observed log2 signal fold change ($log2(STARR_{2i}/STARR_{Serum})$).

3.2.4 DNA methylation analysis

DNA methylation at CpG dinucleotides was investigated using published whole-genome bisulfite sequencing data (WGBS) [91, 278]. Preprocessed WGBS samples were downloaded and listed in Supp. Table B.8. Since the preprocessed data was mapped to *mm9*, we converted our enhancer coordinates from *mm10* to *mm9* using liftOver. CpG positions and methylation levels were loaded into R as a GenomicRange object which was subsequently used to compute average methylation levels as well as methylation metaprofiles with a custom script [191]. Methylation metaprofiles were smoothed using the 100bp running average.

3.2.5 Proteomics analysis

Processed proteomics datasets are listed in Supp. Table B.10. ChIP-SICAP data was already preprocessed and log2 fold-change between ESC conditions ($log2(2i/serum)$) was taken from the published table [204]. The table of ESC whole chromatin binding proteins table was downloaded from supplementary data of [205] but the log2 fold-change between ESC conditions ($log2(2i/serum)$) was missing [205]. We performed a differential protein abundance analysis using the function LFQ from the R-package DEP with the following parameters: alpha=0.05, fun="knn" [279]. Briefly, this function performs missing value imputation using k-nearest neighbour, variance stabilising normalisation and differential enrichment analysis using limma [280, 281]. The resulting differential abundance statistics for chromatin associated proteins were used for downstream analysis.

3.2.6 Prediction of condition-specific enhancers using chromatin features

First, we identified other chromatin associated proteins which might be predictive for condition specific enhancer activity. The initial search space was determined by proteins identified by chromatin proteomics that did not intersect a recent published TF dataset [63] (Supp. Table B.11). ChIP-seq peaks associated with these proteins were taken from Cistrome and for each peak set a differential enrichment analysis using Fisher's exact test was performed (Supp. Table B.11). Resulting $log2(Odds)$ values were plotted against differential protein abundance $log2(2i/serum)$[193]. Factors of interest were determined for each condition by applying the following thresholds for 2i: $log(Odds) > 1 \& log2(2i/serum) > 0$ and for serum: $log2(Odds) < -1 \& log2(2i/serum) < 0$. ChIP-seq data for factors of interest were

downloaded and pre-processed as described above.

Next, we computed a score for each chromatin feature of interest including chromatin accessibility (ATAC-seq), CpG methylation (WGBS), histone modifications and other chromatin associated proteins (ChIP-seq). Depending on the data type or availability, different types of scores were computed (see Figure 3.15 for scores). The resulting scores were used as classifiers to distinguish 2i from serum enhancers and performance was evaluated by computing the area under ROC curve (auROC). Since the auROC is sensitive to unbalanced classes, we performed 100 re-sampling iterations each time sampling 300 examples per class and computing the auROC from them. Finally, auROCs were summarised as mean and standard deviation per classifier.

3.2.7 Prediction of condition-specific enhancers activity using sequence features

Condition-specific enhancers were split into training set and hold-out test sets according to their chromosomal location as follows: training set chr1-chr8, chr10, chr12-15; test set chr9, chr11, chr14-19, chrX, chrY. Chromosome wise splitting was performed following a recent publication suggesting this approach to avoid potential overlaps and overfitting [282]. 2i enhancers and serum enhancers served as target variables for model training.

GBM feature selection, hyperparameter optimisation and model training

Feature selection based on assigning predictive PWMs to proteomics defined TFs was performed as described in Section 2.2.13. Finally, feature sets were compiled by filtering for high predictive power ($auROC > 0.6$) and split into four distinct feature sets: one set for all ESC chromatin associated TFs identified by van Mierlo *et al*, another one for all NANOG co-binding TFs identified by Rafiee *et al*, one representing the overlap between the first two sets ("overlap") and another one combining these two sets ("All") [205, 204].

Subsequently we trained one GBM for each of these four feature sets while performing GBM hyperparameter optimisation as described in Section 2.2.13. One variation was that we used class weights during training due to an imbalanced class distribution. We computed class weights as: $w_y = \frac{N}{2 \cdot n_y}$ where N is the total number of sequences and n_y the number of sequences in class y. Another variation to this procedure was that we tracked the area under

the precision-recall (auPRC) instead of Matthews Correlation Coefficient (MCC) due to the larger class imbalance.

gkm-SVM hyperparameter optimisation and model training

Hyperparameter optimisation as well as gkm-SVM training were performed as described in Section 2.2.13.

Evaluation and comparison of trained models

All trained models were evaluated on the hold-out test set using following performance measures: area under precision recall curve (auPRC), Matthews correlation coefficient (MCC), balanced accuracy, sensitivity and specificity [194]. Since the MCC can be sensitive to unbalanced classes, we performed 100 re-sampling iterations each time sampling 100 examples per class from the test set and computing the MCC from them. Finally, MCCs were summarised as the mean and standard deviation for each model.

GBM feature importance analysis

Scaled feature importance scores for the Nanog co-binding TF GBM model were computed using the function varImp from the R-package caret [206]. varImp is a wrapper function for the relative.influence function from the gbm R-package [207]. Briefly, the relative influence of a feature is defined as loss improvement made in each split the feature was used averaged over all trees [211].

3.3 Results

In this chapter, we investigate the features underlying differential enhancer activities between naive (2i) and primed (serum) ESCs that could explain the observed differences in gene expression and ultimately, phenotypic differences between the cell types. We focused again on NANOG binding sites as putative enhancer elements that potentially regulate differential gene expression. To determine enhancer elements amongst NANOG TFBS and define their differential activity, we used a high-throughput enhancer reporter assay called ChIP STARR-seq was used [178].

3.3.1 Definition of NANOG binding enhancers specific to 2i and serum ESC conditions

First, we used our previously defined high-confidence set of NANOG binding sites as the search space for putative differentially active enhancers (see Section 2.3.1). Enhancer activities of these TFBS were quantified by counting STARR-seq fragments from 2i and serum ESC conditions. These counts were used as the input to an edgeR analysis resulting in 813 enhancers that are statistically significantly higher expressed under 2i conditions and 631 ones under serum conditions ($FDR \leq 0.05$) [195]. Subsequently, NANOG TFBS were filtered for their enrichment over the transfected vector library ($log2(FC) > 0$) to determine the final sets of *bona fide* enhancer elements. These final sets included 768 naive (2i) specific, 481 primed (serum) specific and 6547 non-specific (both) transcriptional enhancers (Fig. 3.2A). This analysis was constrained by the same experimental factors (core-promoter, no UMIs, etc.) as previously discussed in Chapter 2 [176, 219]. Nevertheless, we observed some of the top 300 differentially expressed genes between 2i and serum ESC, as the nearest genes to our differentially active enhancers (Fig. 3.2A). For 2i-specific enhancers, genes included *Prdm14* which was described by several studies as a crucial factor in maintaining a naive pluripotent phenotype [283, 284]. For serum specific enhancers, genes included *Lefty1* which was described in the literature as being involved in differentiation regulation [285]. Moreover, integrative analysis of the STARR-seq differential statistics and the differential RNA-seq revealed a trend between differential enhancer activity and differential expression of the closest gene (Fig. 3.2C). The first two FDR quantiles (most differentially expressed enhancers) showed significant differences in the expression of nearby genes between 2i and serum ($p - value < 0.01$, two-sided Wilcoxon test). Additionally, we measured the closest distance between condition-specific enhancers and differentially expressed genes

(Fig. 3.2D). This revealed that serum enhancers were located in closer proximity to serum differential genes compared with 2i enhancers ($median_{2i} = 5.81$, $median_{serum} = 5.68$; p-value=$4.2 \cdot 10^{-5}$, two-sided Wilcoxon test). However, a similar trend was not observable for 2i specific enhancers and 2i differential genes ($median_{2i} = 5.74$, $median_{serum} = 5.72$; p-value=0.31, two-sided Wilcoxon test).

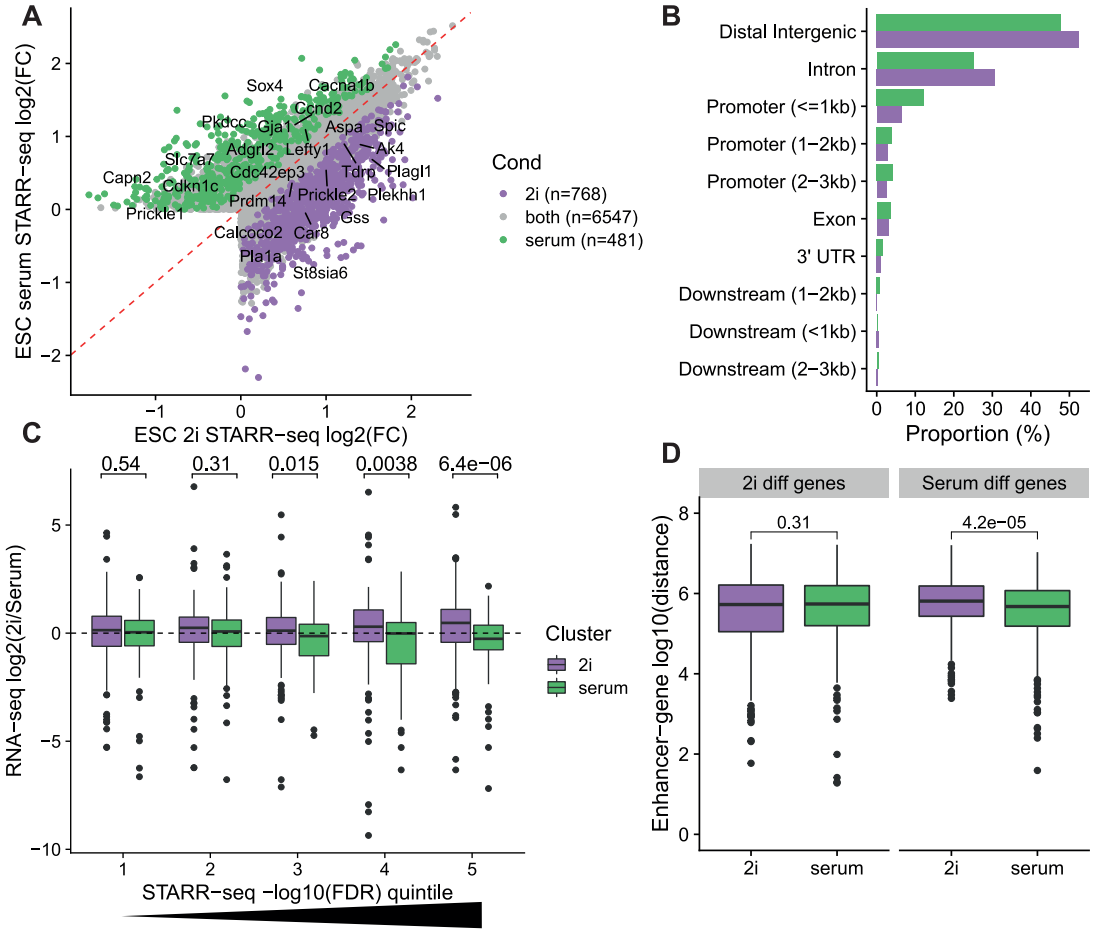


Fig. 3.2: Definition of naive (2i) and primed (serum) NANOG binding pluripotency enhancers. **(A)** Scatterplot comparison of enhancer activity, measured as fold-change over vector library, in naive (2i) against primed (serum) mESC conditions. Enhancers were categorised into the following three categories: naive (2i) specific (purple), primed (serum) specific (green) and active in both conditions (grey). Differentially expressed genes between 2i and serum mESC that were proximal to NANOG enhancers are labelled. Differentially expressed genes were defined as top 300 smallest FDRs. **(B)** Proportion of genome annotations overlapping 2i and serum enhancers. Naive-specific shown in purple and primed-specific coloured in green as in (A). **(C)** Boxplot illustrating the relationship between STARR-seq (-log10(FDR)) and differential expression (log2(FC)) of the nearest genes. P-values were computed using a two-sided Wilcoxon test. **(D)** Boxplot comparison of enhancer-gene distances for differentially expressed genes between 2i (purple) and serum (green) specific enhancers. P-values were computed using a two-sided Wilcoxon test.

In terms of genome-wide distribution, condition-specific enhancers were mostly located in distal intergenic (2i: 52.5%, 403/768; serum: 47.8%, 230/481) or intronic regions (2i: 30.7%, 236/768; serum: 25.1%, 121/481; Fig. 3.2B). Serum-specific enhancers showed a higher proportion of promoter overlap compared with 2i enhancers (2i: 6.4%, 49/768; serum: 12.3%, 59/481).

Having defined 2i- and serum-specific transcriptional enhancers, we selected several examples for locus-specific examination. For 2i enhancers, we focused on the differentially active enhancer that might be associated with *Prdm14* (Fig. 3.3). Examination of the condition specific NANOG, OCT4, SOX2 ChIP-seq revealed binding of these factors to similar extents independent of profiled condition. The ChIP-seq measurements of the primed enhancer mark H3K4me1 and activating mark H3K27ac did not show condition specific differences either. However, open-chromatin profiling by ATAC-seq revealed substantial differences in chromatin accessibility within the enhancer element between 2i and serum ESC condition (Fig. 3.3). This might explain the observed differential activity between both ESC conditions. Interestingly, we also observed an enrichment of the repressive Polycomb mark H3K27me3 at the *Prdm14* promoter in serum ESC.

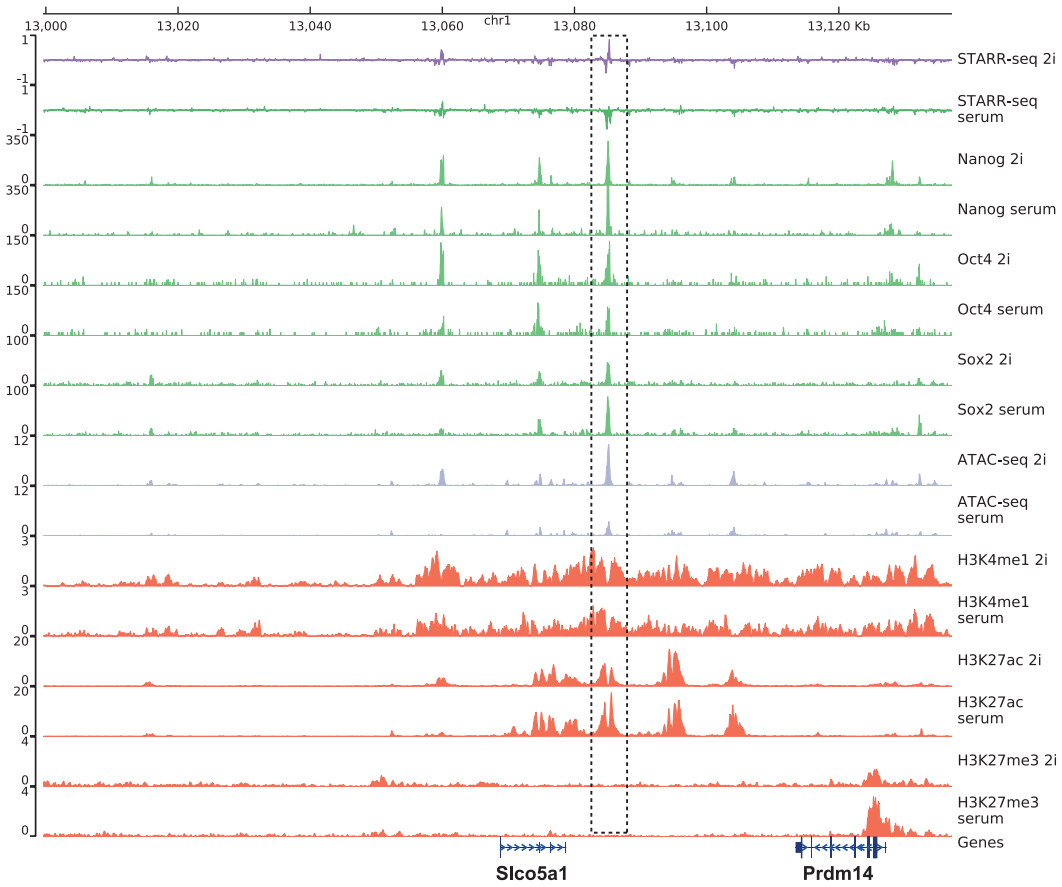


Fig. 3.3: Example for naive (2i) mESC specific transcriptional enhancer at the *Prdm14* locus. Naive(2i) mESC specific enhancer element potentially involved in driving expression of *Prdm14* was highlighted. Signal tracks show (from top to bottom): STARR-seq 2i log2(FC) (purple); STARR-seq serum log2(FC) (green); TF ChIP-seq for NANOG, OCT4 and SOX2 form 2i and serum mESCs as background subtracted signals (green); ATAC-seq signal from 2i and serum ESCs as CPMs (purple); histone PTM ChIP-seq signal for H3K4me1, H3K27ac and H3K27me3 from 2i and serum mESCs normalised to CPM (red).

Next, as an example of a serum-specific enhancer we examined an element located between *Arl4d* and *Dhx8* (Fig. 3.4). TF ChIP-seq showed again binding of NANOG, OCT4 and SOX2 in both conditions, but SOX2 showed higher binding in serum. Moreover, H3K4me1 ChIP-seq exhibited similarly high enrichment levels in both conditions and H3K27me3 was depleted on this. In contrast to the 2i enhancers above, this enhancer showed only low levels of chromatin accessibility as measured by ATAC-seq. Nevertheless, we observed differential ChIP-seq enrichment of the activating histone mark H3K27ac between 2i and serum cultured ESCs (Fig. 3.4). The acquisition of H3K27ac upon culture of ESC in serum medium is in support of the differential activity of this specific element.

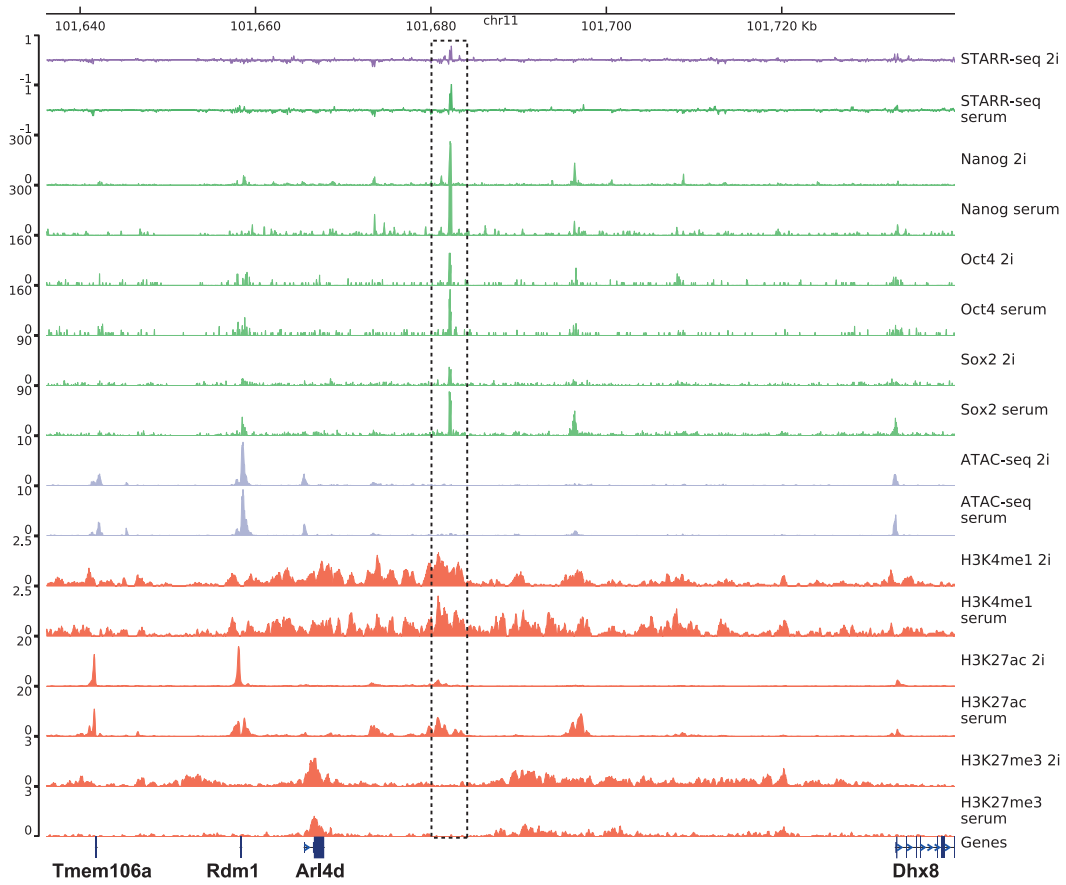


Fig. 3.4: Example for primed (serum) mESC specific transcriptional enhancer at the *Arl4d* locus. Primed (serum) mESC specific enhancer element with unknown target gene was highlighted. Signal tracks show (from top to bottom): STARR-seq 2i log2(FC) (purple); STARR-seq serum log2(FC) (green); TF ChIP-seq for NANOG, OCT4 and SOX2 form 2i and serum mESCs as background subtracted signals (green); ATAC-seq signal from 2i and serum ESCs as CPMs (purple); histone PTM ChIP-seq signal for H3K4me1, H3K27ac and H3K27me3 from 2i and serum mESCs normalised to CPM (red).

In this section we defined transcriptional enhancers exhibiting statistically significant differential activity between naive (2i) and primed (serum) ESCs. These enhancers showed proximity to differentially expressed genes and as expected most of them were located in intergenic or intronic regions. Examination of specific examples suggested possible determinants of differential enhancer activity.

3.3.2 Binding of pluripotency TFs at naive (2i) and primed (serum) ESC enhancers

A crucial hallmark of transcriptional enhancers, as previously mentioned, is TF binding and in case of pluripotency enhancers especially binding of NANOG, OCT4 and SOX2 [286]. Hence, we investigated the binding dynamics of these three factors to our condition-specific enhancers. NANOG ChIP-seq generated for this study revealed similar average binding signals for 2i and serum enhancers with ChIP-seq profiling under ESC 2i conditions (Fig. 3.5A). No significant differences were observable in NANOG 2i ChIP-seq signals between the two sets of enhancers ($median_{2i} = 1.4$, $median_{serum} = 1.78$, $p = 0.49$, one-sided Wilcoxon test; Fig. 3.5B). In contrast, assaying NANOG binding under ESC serum conditions showed higher average binding of NANOG to serum enhancers compared with 2i enhancers (Fig. 3.5A). This observation was further supported by enrichment of NANOG serum ChIP-seq signal within serum enhancers compared with 2i enhancers ($median_{2i} = 1.65$, $median_{serum} = 2.12$, $p = 3.4 \cdot 10^{-7}$, one-sided Wilcoxon test; Fig. 3.5B).

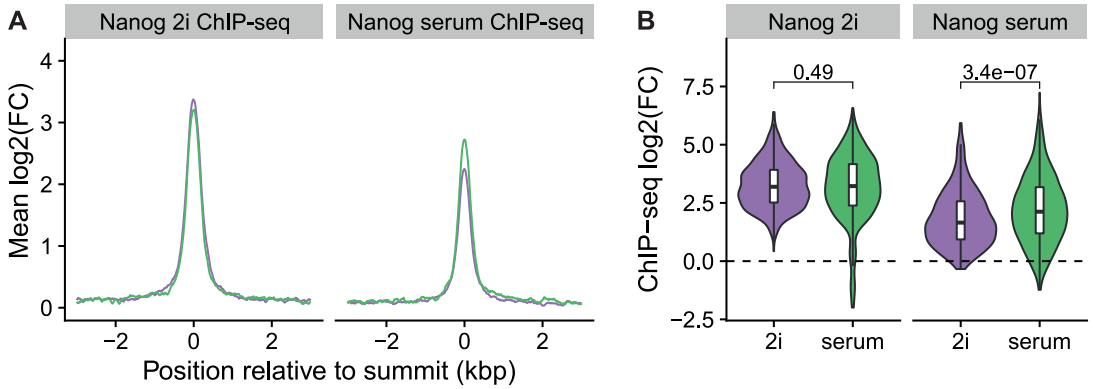


Fig. 3.5: NANOG binding to naive (2i) and primed (serum) enhancers. (A) Metaprofile showing average NANOG ChIP-seq signal ±3kb around 2i (purple) and serum (green) enhancer summits. ChIP-seq was performed for 2i (left panel) and serum (right panel) mESC condition and shown as normalised as log2 fold change over input control. **(B)** Violin plots comparing NANOG ChIP-seq signal within 2i and serum enhancers. The signal was computed within ±250bp around peak summits. ChIP-seq was performed for 2i conditions (left panel) and serum mESC conditions (right panel). P-values derived from one-sided Wilcoxon test.

Next, we made use of previously published ChIP-seq samples to investigate the binding dynamics of all three TFs in three ESC conditions including: 24h 2i, three passages 2i (3p2i) and serum [216]. We observed on average binding of all three factors in all available conditions to both enhancer types (Fig. 3.6A).

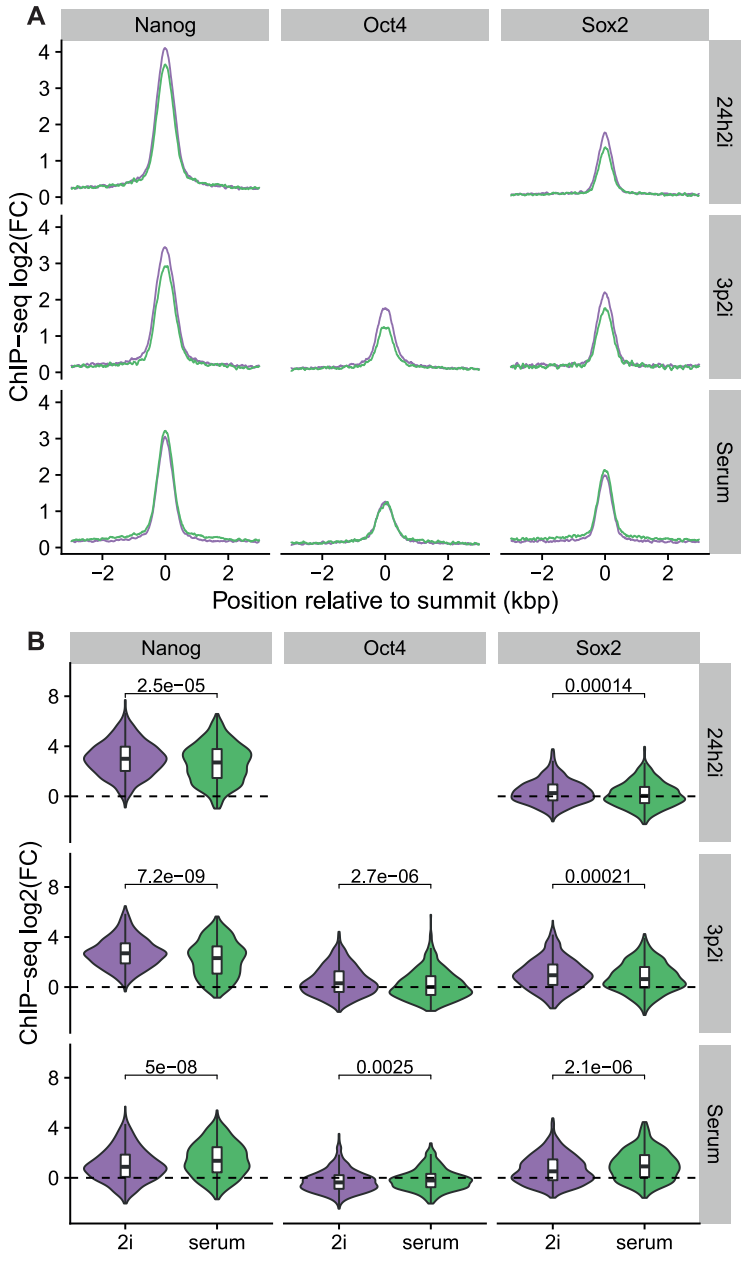


Fig. 3.6: Nanog binding to naive (2i) and primed (serum) enhancers. (A) Metaprofiles showing average NANOG, OCT4 and SOX2 ChIP-seq signal (log2 fold-change over input) around 2i (purple) and serum (green) enhancer summits. ChIP-seq samples were previously published and included NANOG (left panel), OCT4 (middle panel) and SOX2 (right panel) for ESC conditions 24h 2i (top panel), three passages 2i (3p2i) (middle panel) and serum (bottom panel) [216]. **(B)** Violin plot comparisons of NANOG, OCT4 and SOX2 ChIP-seq signals within 2i (purple) and serum (green) enhancers. P-values were computed using one-sided Wilcoxon test.

Nevertheless, 2i-specific enhancers exhibited higher average binding of NANOG, OCT4 and SOX2 in ChIP-seq from ESC 2i than serum enhancers. In contrast, serum enhancer showed slightly higher average binding of NANOG/SOX2 and equally high binding of OCT4 under ESC serum conditions compared with 2i enhancers (Fig. 3.6A). Quantification of ChIP-seq enrichment for each enhancer element revealed, as expected, higher levels of TF binding to enhancers within their respective condition. For example, NANOG showed significantly higher binding to 2i enhancers in 3p2i conditions and serum enhancers significant higher engagement under serum conditions (3p2i: $median_{2i} = 2.7$, $median_{serum} = 2.23$, $p = 7.2 \cdot 10^{-9}$, serum: $median_{2i} = 0.88$, $median_{serum} = 1.36$, $p = 5 \cdot 10^{-8}$, one-sided Wilcoxon test). Long term culture of ESC under 2i conditions increases binding of SOX2 to enhancer elements compared with only 24h culture (24h2i: $median_{2i} = 0.26$, $median_{serum} = 0.03$; 3p2i: $median_{2i} = 0.95$, $median_{serum} = 0.64$; serum: $median_{2i} = 0.52$, $medianserum = 0.91$; Fig. 3.6B). In contrast, OCT4 showed binding only to a subpopulation of enhancers for ChIP-seq from 3p2i as well as serum ESCs (Fig. 3.6B). Nevertheless, OCT4 enrichment was condition-specific, as mentioned above, meaning that 2i enhancers exhibited significantly higher OCT4 levels under 3p2i conditions and vice versa (3p2i: $median_{2i} = 0.3$, $median_{serum} = 0.01$, $p = 2.7 \cdot 10^{-6}$; serum: $median_{2i} = -0.37$, $median_{serum} = -0.21$, $p = 0.0025$, one-sided Wilcoxon test).
In summary, NANOG, OCT4 and SOX2 binding, as measured by ChIP-seq, at 2i- and serum-specific enhancers tended to show greater occupancies in their respective conditions: 2i enhancers showed higher TF levels if profiled under 2i conditions and vice versa. These findings provide evidence for association with differential enhancer activities, though not causality.

3.3.3 Differences in the epigenetic landscape around naive (2i) and primed (serum) ESC enhancers

Another important enhancer hallmark is the enrichment of specific histone marks such as H3K4me1 for primed enhancers and H3K27ac for active enhancers [83]. To find further support for the condition-specific activity of our enhancer sets, we examined ChIP-seq enrichment of these two histone marks, as well as repressive H3K27me3 and histone acetylase P300 [231]. The average ChIP-seq signals reflected depleted nucleosome occupancies within enhancer regions corresponding to nucleosome free regions (NFRs) and binding of P300 to these NFRs (Fig. 3.7A). H3K4me1 showed little differences on average between 2i and serum enhancers in ChIP-seq measurements in their respective conditions. Moreover, no significant changes could be found in ChIP-seq enrichments from 2i and serum ESC cells

within both enhancer classes ($median_{2i} = -0.01$, $median_{serum} = 0.03$, $p = 0.83$, two-sided Wilcoxon test; Fig. 3.7B). Enhancer activity changes were not correlated with ChIP-seq enrichment changes at 2i enhancers and displayed only weak correlation at serum enhancers (2i: $R = -0.021$, serum: $R = 0.15$; Fig. 3.7C). In contrast, the activating histone mark H3K27ac showed on average condition-specific ChIP enrichment for 2i enhancers but only small differences for serum enhancers (Fig. 3.7A). Quantification of ChIP enrichment changes between conditions within these enhancer elements revealed significant differences between them and signal changes for each enhancer set towards their respective conditions ($median_{2i} = 0.21$, $median_{serum} = -0.67$, $p < 2.2 \cdot 10^{-16}$, two-sided Wilcoxon test; Fig. 3.7B). Nevertheless, we could only find a weak correlation of STARR-seq changes and ChIP-seq changes within 2i enhancers and no correlation for serum enhancers (2i $R = 0.17$, serum $R = 0.032$; Fig. 3.7C). We investigated P300, the histone acetylase that establishes H3K27ac, which again showed condition-specific enrichment on average for both enhancer classes, although observed differences in average signal were larger for ChIP-seq from ESC serum conditions (Fig. 3.7A). P300 exhibited significant ChIP-seq enrichment changes between ESC cells cultured under 2i or serum conditions ($median_{2i} = 0.91$, $median_{serum} = -0.42$, $p < 2.2 \cdot 10^{-16}$, two-sided Wilcoxon test; Fig. 3.7B). These signal changes were towards their expected direction and showed largest difference for all examined ChIP-seq samples in this subsection. Nevertheless, we could not observe any correlation between STARR-seq and P300 changes for serum enhancers and only a weak one for 2i enhancers (2i $R = 0.12$, serum $R = 0.075$; Fig. 3.7C). Lastly, we examined the repressive Polycomb mark H3K27me3 which showed higher average signal on serum enhancers regardless of the profiled ESC condition (Fig. 3.7). Additionally, no significant differences in H3K27me3 ChIP enrichment was observable ($median_{2i} = -0.26$, $median_{serum} = -0.24$, $p = 0.016$, two-sided Wilcoxon test; Fig. 3.7B). However, H3K27me3 ChIP-seq from serum ESC showed higher signal levels at both enhancer classes. Also, we could not observe any correlation between STARR-seq and H3K27me3 (2i: $R = 0.042$, serum: $R = -0.016$; Fig. 3.7C).

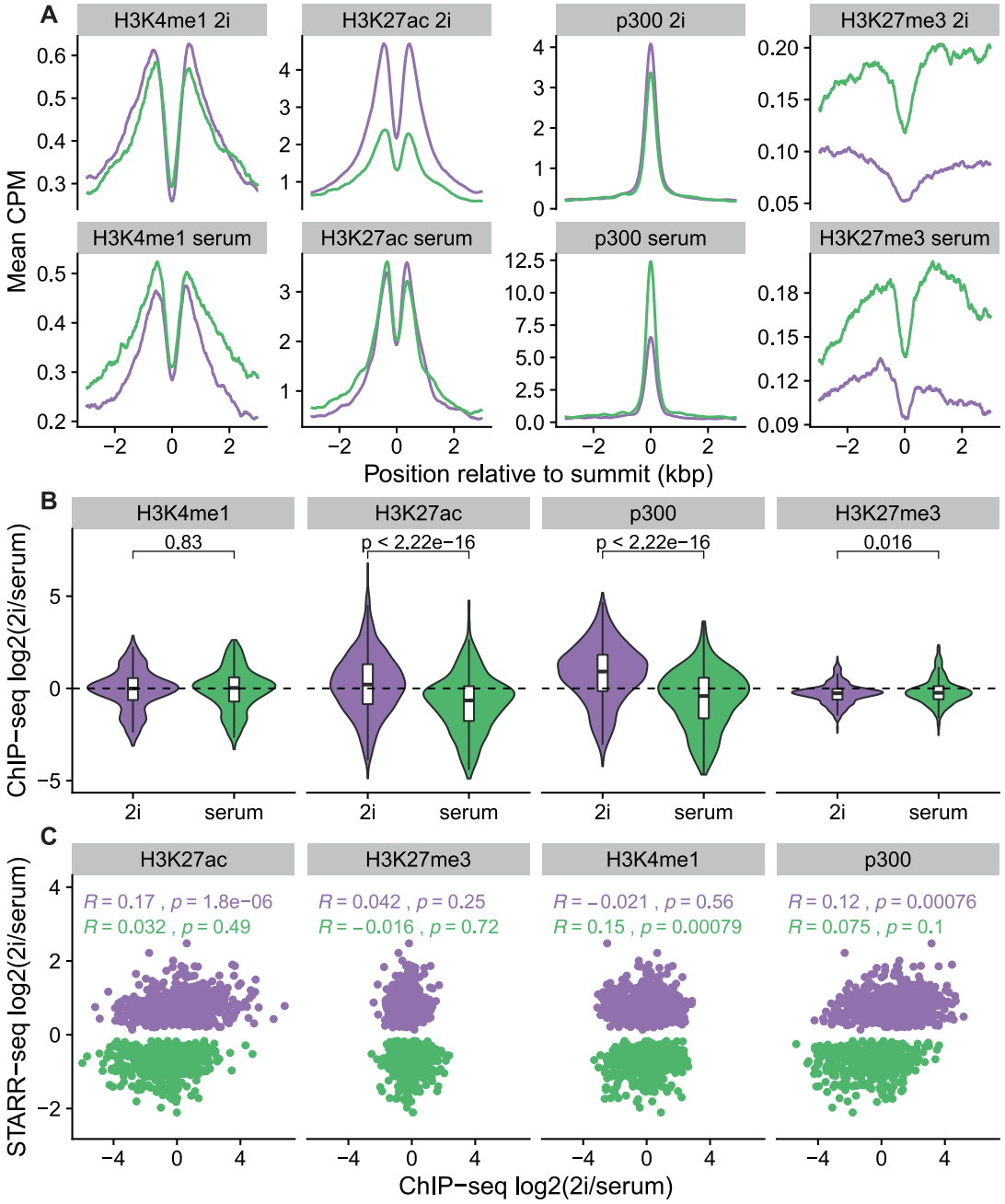


Fig. 3.7: Epigenetic landscape of naive (2i) and primed (serum) enhancers. **(A)** Metaprofiles showing average CPM ChIP-seq signals for $\pm 3kb$ around 2i (purple) and serum (green) enhancer summits. ChIP-seq samples included H3K4me1, H3K27ac, P300 and H3K27me3 for 2i (top) and serum (bottom) mESC conditions [231]. **(B)** Comparison of ChIP-seq signal changes between 2i and serum mESC conditions for H3K4me1, H3K27ac, P300 and H3K27me3 within 2i (purple) and serum (green) specific enhancers. P-values were computed using a two-sided Wilcoxon test. **(C)** Scatterplot comparison of STARR-seq log2 fold-change and ChIP-seq log2 fold-change between 2i and serum mESC conditions for both enhancer classes. Pearson correlation and p-value derived from correlation test.

Altogether, we confirmed ChIP-seq enrichment of enhancer relevant histone marks such as H3K4me1 and H3K27ac as well as the histone acetylase P300 on enhancer peaks. Moreover, we found additional support for the differential activity of our condition-specific enhancers by condition-specific enrichment of P300 as well as H3K27ac.

3.3.4 Open chromatin dynamics of naive (2i) and primed (serum) ESC enhancers

Another important enhancer hallmark is chromatin accessibility, meaning nucleosome depletion, due to TF engagement and activating histone modifications. Therefore, we investigated chromatin accessibility dynamics of our condition-specific enhancer elements using ATAC-seq from different timepoints of the respective ESC culture conditions [199]. First, ATAC-seq from ESCs cultured for 8 hours in either 2i or serum already started to exhibit condition-specific accessibility for both enhancer classes (Fig. 3.8 first panel). However, these differences were small which was most likely explainable by cell-to-cell heterogeneity in the unsynchronised cell culture. After one day significant higher chromatin accessibilities were observable for enhancers in their respective conditions (2i $p = 3.7 \cdot 10^{-}10$, serum $8.5 \cdot 10^{-6}$, one-sided Wilcoxon test; Fig. 3.8 second panel). This accessibility trend continued in the day three and long-term (LT) cultures, revealing both further increases in enhancer accessibility in the respective conditions. We observed in 2i LT cultured ESC significant higher accessibilities within 2i enhancer elements and vice versa for serum enhancers (2i $p < 2.22 \cdot 10^{-16}$, serum $p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 3.8, fourth panel). These observations might be explainable with increasing adaptation of ESCs to their culture conditions over time and therefore ATAC-seq, as a bulk measure, exhibiting larger differences under longer culture times. Taken together, ESC condition-specific enhancer elements displayed higher levels of accessibility in their respective condition and therefore further supporting their specific activity in 2i/serum condition.

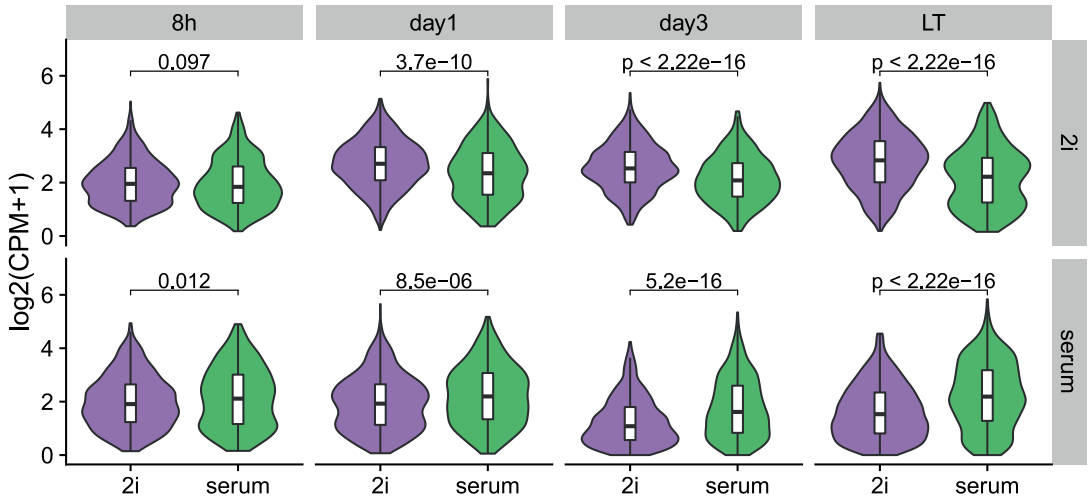


Fig. 3.8: Open chromatin dynamics of naive (2i) and primed (serum) enhancers. Comparison of chromatin accessibility, as measured by ATAC-seq, within 2i (purple) and serum (green) specific enhancers. ATAC-seq samples included two different conditions 2i (top) and serum (bottom) at four different time-points of culture 8h, one day, three days and long-term (LT) [199]. ATAC-seq signal was computed as log2(CPM+1) and shown as violin plot. P-values were computed using one-sided Wilcoxon test.

3.3.5 DNA methylation dynamics at (2i) and primed (serum) ESC enhancers

ESC cultured under 2i conditions exhibited genome-wide hypomethylation which is in line with showing a more ground-state/naive pluripotency phenotype. Changing the culture condition of these ESCs to serum induced establishment of DNA methylation particularly at CpG dinucleotides [91]. To investigate how DNA methylation differences might impact enhancer activities, we analysed previously published whole-genome bisulfite sequencing (WGBS) [91, 278]. We observed on average low levels of CpG methylation within our enhancer elements under ESC 2i conditions and there was no significant difference in methylation levels between enhancers ($median_{2i} = 2.9$, $median_{serum} = 3.57$, $p = 0.046$, one-sided Wilcoxon test; Fig. 3.9A, B first panel). Transfer from 2i to serum conditions set up CpG methylation as previously described but both enhancer types stayed lowly methylated (Fig.3.9A, B second panel) [278]. However, 2i specific enhancers seemed to acquire more DNA methylation and serum enhancers displayed significantly lower methylation under these conditions ($median_{2i} = 18.3$, $median_{serum} = 10$, $p = 8.8 \cdot 10^{-7}$, one-sided Wilcoxon test). WGBS from long-term serum cultured ESCs revealed on average low levels of methylation within serum enhancers and further gain on 2i enhancers (Fig. 3.9A third panel).

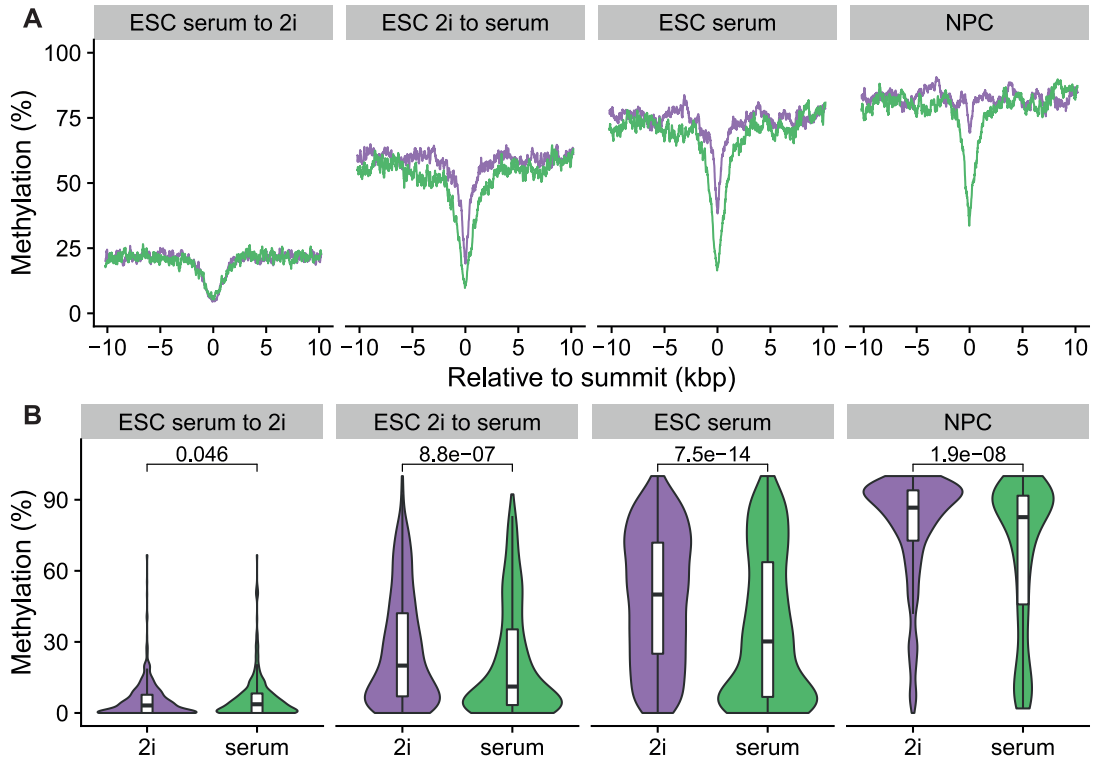


Fig. 3.9: Methylation levels at naive (2i) and primed (serum) enhancers. (A) Metaprofiles summarising average methylation levels, as measured by whole-genome bisulfite sequencing (WGBS), for $\pm 5kb$ around 2i (purple) and serum (green) enhancer summits. Methylation profiles were smoothed using 100bp running average. WGBS samples included mESC transferred from serum to 2i, from 2i to serum, long-term serum culture and differentiated to neural progenitor cells (NPC) [91, 278]. **(B)** Violin plot comparison of mean methylation levels within 2i (purple) and serum (green) enhancers. P-values were derived from one-sided Wilcoxon test.

Additionally, most 2i enhancers continued to gain CpG methylation under long term culture conditions and serum enhancers remained significantly lowly methylated ($median_{2i} = 48$, $median_{serum} = 29.2$, $p = 7.5 \cdot 10^{-14}$, one-sided Wilcoxon test; Fig. 3.9B third panel). Finally, differentiation of ESCs into neural progenitor cells was accompanied by further increases in DNA methylation levels at both enhancer classes (Fig. 3.9A fourth panel). Almost all 2i enhancers now exhibited high methylation levels ($> 90\%$) as well as a majority of serum enhancers. Although, serum enhancers all together still showed significantly lower methylation levels than 2i enhancers ($median_{2i} = 85.9$, $median_{serum} = 81.5$, $p = 1.9 \cdot 10^{-8}$, one-sided Wilcoxon test). In summary, we observed faster gain of repressive CpG methylation at 2i enhancer compared with serum enhancers upon ESC removal from 2i conditions and

almost full methylation upon differentiation. This supports the specificity of these condition-specific elements for their respective ESC conditions.

3.3.6 Nascent eRNA transcription at (2i) and primed (serum) ESC enhancers

A hallmark of active enhancer elements is exhibiting bi-directional nascent transcription of eRNAs which can be measured with sequencing techniques such as GRO-seq [159, 161]. To provide additional evidence for the differential activity of our enhancer elements we analysed GRO-seq data from ESCs cultured in 2i as well as serum [247]. This revealed bi-directional expression of eRNAs from 2i enhancers and slightly unbalanced transcription from serum enhancers as measured by GRO-seq from 2i ESCs (Fig. 3.10A, left panel). 2i enhancers showed significantly higher eRNA transcription in this GRO-seq condition than serum enhancers ($p = 8.6 \cdot 10^{-12}$, one-sided Wilcoxon test; Fig. 3.10B, left panel). In contrast, GRO-seq from serum ESCs showed little if any transcription at 2i enhancers and weak bi-directional expression from serum enhancers (Fig. 3.10A, right panel). Quantification of nascent transcription confirmed that most 2i enhancers were not transcribed and serum enhancers showed significantly higher expression levels ($p = 1.3 \cdot 10^{-10}$, one-sided Wilcoxon test; Fig. 3.10B, left panel). A possible explanation for this stark contrast might be the acquisition of repressive DNA methylation during the transition from 2i to serum [287]. Altogether, we found additional evidence that our set of 2i and serum enhancers correspond to *bona fide* enhancer elements and that these enhancers exhibit condition-specific properties as anticipated by their definition.

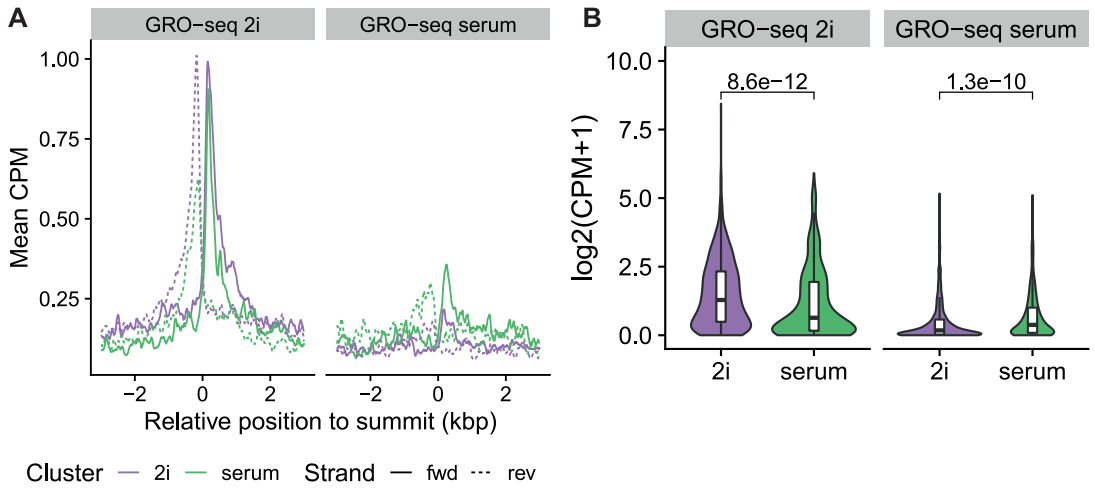


Fig. 3.10: Nascent eRNA transcription at naive (2i) and primed (serum) enhancers. (A) Metaprofile summarising average nascent eRNA expression, as measured by GRO-seq, within $\pm 3kb$ around 2i (purple) and serum (green) enhancer summits. eRNA Expression was normalised to CPM. GRO-seq samples included mESC 2i (left panel) and serum condition (right panel) [247]. Linetype indicates the strand reads derive from, solid = forwards and dotted = reverse. **(B)** Violin comparisons of nascent eRNA transcription within 2i (purple) and serum (green) enhancers for GRO-seq from 2i (left) and serum conditions (right). GRO-seq signal was summarised for ± 250 around summits. p-values were computed using a one-sided Wilcoxon test.

3.3.7 Repetitive element overlap with naive (2i) and primed (serum) ESC enhancers

Repetitive sequence elements comprise a potential source for regulatory elements (reviewed in [42]). Previous studies showed that these elements are highly enriched for TF binding sites and that regulatory elements in turn were enriched for repeats [288, 289]. For these reasons, we investigated if our 2i and serum enhancers are enriched for sequence repeats by overlapping them with the RepeatMasker annotation. We observed significant enrichment for 14 and 13 out of the 24 RepeatMasker repeat family classes for 2i and serum enhancers respectively ($FDR < 0.01$, one-sided Binomial test; Fig. 3.11 A). These repeat families included long terminal repeats (LTRs) e.g. ERVK (2i: 18.9% (145/768), serum: 17.3% (83/481)), short interspersed nuclear element (SINE) e.g. Alu (2i: 11.7% (90/768), serum: 17.3% (83/481)), simple repeats (2i: 10% (77/768), serum: 14.1% (68/481)) and low complexity repeats (2i: 2.99% (23/768), serum: 7.28% (35/481)). These findings are in line with recent reports of LTRs such as ERVK elements and SINEs displaying enhancer function in ESCs and other cell types [290, 291, 292]. However, differential repeat enrichment analysis between 2i and serum enhancer elements revealed no significant differences between both enhancer classes

(Fig. 3.11 B). Taken together, we recovered enrichment of repetitive elements such as LINEs and SINEs known to function as enhancer elements in ESC for both enhancer sets.

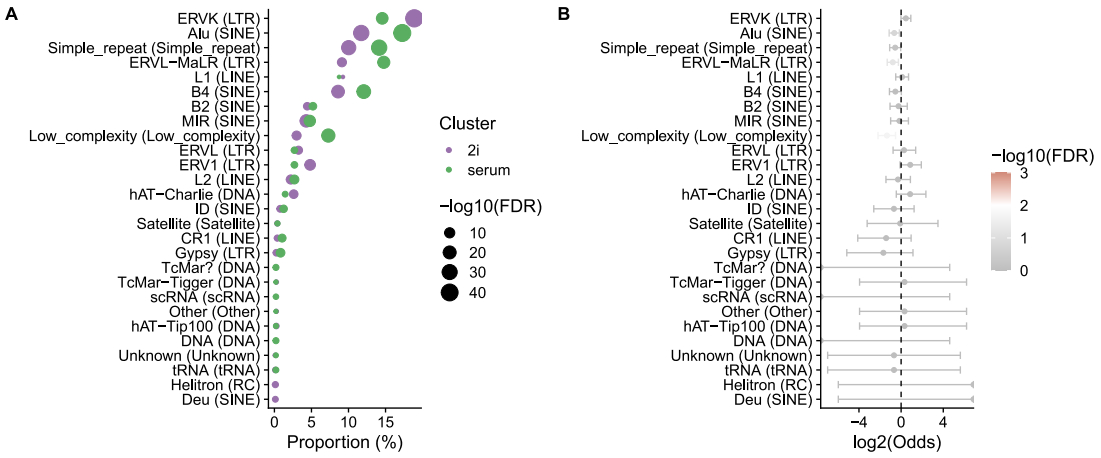


Fig. 3.11: Enrichment of sequence repeats in naive (2i) and primed (serum) enhancer sets. (A) Proportion of 2i (purple) and serum (green) enhancers overlap with RepeatMasker repetitive element families. FDRs were computed using a one-sided Binomial test followed by multiple testing correction with the Benjamini-Hochberg procedure. **(B)** Differential enrichment of RepeatMasker repetitive element families between 2i and serum enhancers. FDRs were computed using a two-sided Fisher's exact test followed by multiple testing correction with the Benjamini-Hochberg procedure

3.3.8 Chromatin-chromatin interactions of naive (2i) and primed (serum) ESC enhancers

Enhancers can influence the expression of their target genes by physically interacting with associated promoter regions [141]. Thus, we investigated how chromatin-chromatin interactions of our condition-specific enhancer would change in different ESC conditions. We first overlapped our enhancer sets with a previously published set of high-confidence interactions, from Atlasi *et al* based on capture Hi-C from 2i and serum ESCs [199]. This analysis revealed 1,029 and 661 interactions in which our 2i and serum enhancers were involved, respectively. Interactions of 2i enhancers were distributed as follows: 14.2% (146/1,029) with promoters, 1.3% (13/1,029) with other 2i enhancers, 1.4% (14/1,029) with serum enhancers and 83.2% (856/1,029) with other genomic elements. Serum enhancer interactions exhibited a similar pattern: 14.4% (95/661) with promoters, 0.5% (3/661) with other serum enhancers, 2.1% (14/661) with 2i enhancers and 83.1% (549/661) interacted with other genomic elements. Next, we compared the capture Hi-C coverage of enhancer interactions between ESC 2i and serum conditions. This revealed that most interactions

remained unchanged between both conditions and showed high correlation between each other, independent of condition-specific enhancer activity (2i $R = 0.88$, serum $R = 0.87$; Fig. 3.12 A). Comparison of log2 coverage fold-changes between 2i and serum condition showed that most interactions showed less than 2-fold changes and only few up to a maximum 4-fold ($median_{2i} = 0.06$, $median_{serum} = -0.004$, $p = 9.1 \cdot 10^{-6}$, two-sided Wilcoxon test; Fig. 3.12 B). These observations were in line with previous findings that there are only a few changes in chromatin-chromatin interactions between these conditions [145, 199]. Comparison of interaction distances between both enhancer classes revealed significant differences between 2i and serum enhancers ($median_{2i} = 4.58$; $median_{serum} = 4.73$; $p = 9.1 \cdot 10^{-6}$, two-sided Wilcoxon test). Serum enhancers were further from their target regions than 2i enhancers (Fig. 3.12C). Taken together, we observed similar interaction trends for our condition-specific enhancers as previously described in the literature [199]. This provided additional evidence that observable gene expression changes between 2i and serum were most likely explained due to differential enhancer activity and not differential chromatin-chromatin interactions.

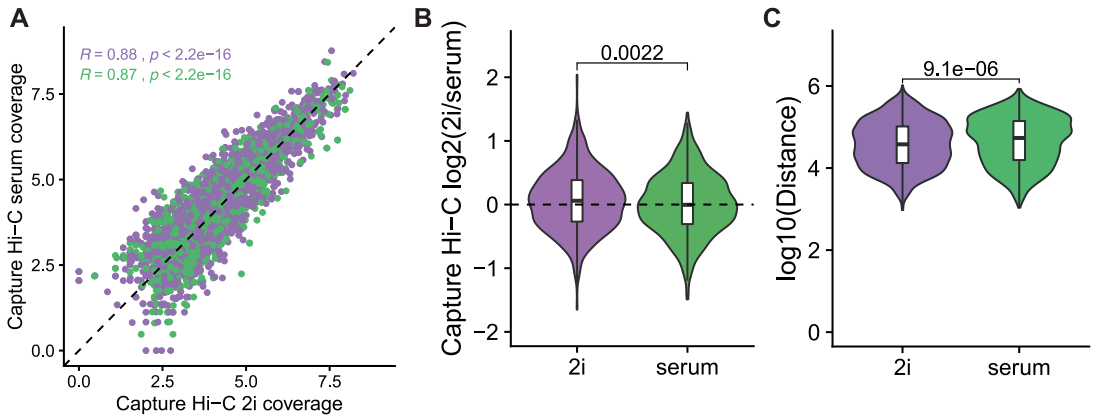


Fig. 3.12: Chromatin-chromatin interactions of naive (2i) and primed (serum) enhancers. (A) Scatterplot comparison of interactions overlapping 2i (purple) or serum (green) enhancers between capture Hi-C samples from mESC 2i and serum conditions [199]. Pearson correlation and p-value were computed using correlation test. **(B)** Violin plot comparisons of capture Hi-C coverage log2 fold-changes between 2i and serum condition for interactions overlapping 2i (purple) or serum (green) enhancers. P-value from two-sided Wilcoxon test. **(C)** Violin plot comparisons of distances for interactions which involve 2i (purple) or serum (green) enhancers. P-value from two-sided Wilcoxon test.

3.3.9 Gene expression of naive (2i) and primed (serum) enhancers target genes

Next, we assigned 2i and serum enhancers to their potential target genes using capture Hi-C interactions from two previously published studies to better understand the impact of condition-specific enhancer on endogenous gene expression and target gene function [199, 146]. This revealed a set of 252 2i and 192 serum target genes using interactions from *Atlasi et al* and 108 2i and 89 serum target genes for Sahlen *et al* [199, 146]. Overlaying these target gene sets with gene expression data from RNA-seq showed that only 3.2% (8/252) of 2i and 3.1% (6/192) of serum target genes, defined with Atlasi *et al* interactions, were not expressed. For Sahlen *et al*, we observed a higher proportion of unexpressed targets genes: 20.4% (22/108) of 2i and 27% (24/89) of serum targets. Nevertheless, comparison of gene expression fold-changes revealed significant differences between 2i and serum targets (Atlasi: $p = 2.8 \cdot 10^{-4}$, Sahlen: $p = 0.012$, two-sided Wilcoxon test; Fig.3.13A). Gene expression differences were especially pronounced for 2i enhancer target genes regardless of the interaction set used to assign them. In contrast, target genes of serum enhancers did not show the expected differences and were on average not differentially expressed (Fig.3.13A). To gain a better understanding of how these potential target genes of condition-specific enhancers might contribute to phenotypic differences between ESCs cultured in 2i vs. serum, we performed a pathway enrichment analysis [201]. This revealed significant enrichment of genes associated with the pluripotency network ("PluriNetWork") in all four different target gene sets (Fig. 3.13B). Moreover, terms such as "ESC Pluripotency Pathways" and "Wnt Signalling Pathway and Pluripotency" were enriched for 2i enhancer target genes (Sahlen) [146]. This is in line with the literature showing the importance of WNT signalling for naive pluripotency [225]. In summary, this analysis of potential target genes is consistent with the importance of our enhancer sets in condition-specific gene regulation and for the involvement of target genes in the different ESC properties changing due to culture.

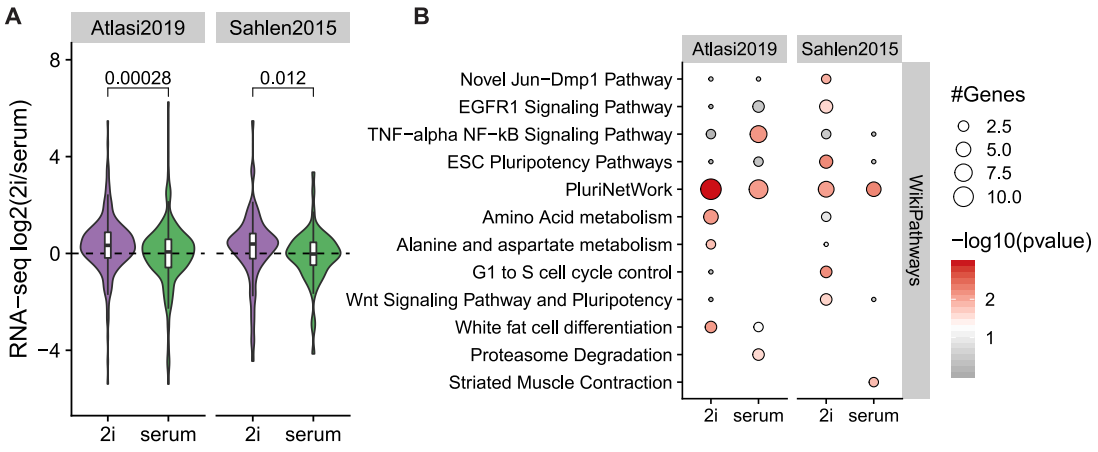


Fig. 3.13: Enhancer target gene expression from 2i/serum RNA-seq and pathway enrichment. **(A)** Comparison of target gene expression changes between 2i and serum mESC conditions. Target genes were assigned to 2i (purple) and serum (green) enhancers using capture Hi-C interactions from two different studies Atlasi *et al* (left) and Sahlen *et al* (right) [199, 146]. P-values were computed using two-sided Wilcoxon test. **(B)** Pathway enrichment analysis of target genes in WikiPathways [201]. Dot size corresponds to the number of pathway genes overlapping target genes. Colour scaled encodes $-log10(p-value)$ from a one-sided Fisher's exact test.

3.3.10 Prediction of condition-specific enhancer activity using chromatin features

After descriptive investigation of the chromatin landscape around 2i- and serum-specific enhancers, we asked which of these features including open chromatin, histone modifications, nascent transcription, methylation and other chromatin-associated factors might be the most predictive for condition-specific enhancer activity. First, we sought to complete the list of potentially predictive chromatin associated proteins by combining data from Cistrome, a publicly available ChIP-seq peak database, with proteomic experiments profiling changes in chromatin-associated proteins between 2i and serum ESC conditions [193]. These two proteomic experiments included profiling all chromatin-bound proteins (Chromatome) and profiling only NANOG co-binding proteins [204, 205]. Combining these datasets revealed that the histone acetylase P300 exhibited higher chromatin abundance under ESC 2i conditions in both experiments, which was confirmed by ChIP-seq peak enrichment (Fig. 3.14A, B). However, our previous analysis of P300 ChIP-seq showed condition-specific signal enrichment on 2i and serum enhancers (Fig. 3.7A, B). Therefore, this analysis strongly depends on the ESC condition in which the ChIP-seq profiling was done; therefore ChIP-seq experimental conditions always were always double checked. Our analysis revealed higher

2i abundance as well as ChIP-seq peak enrichment for the histone methyltransferase ASH2L and the mediator complex subunit MED1 for the ESC Chromatome experiment (Fig. 3.14A). For the Nanog co-binding profiling, we observed high 2i abundance and peak enrichment for CTNNB1, which is in line with previous reports about the contribution of WNT signalling to naive pluripotency [225, 183]. In contrast, we found UTF1 and JARID2 in the fraction of serum-specific chromatin binding proteins, independent of the proteomics study (Fig. 3.14A, B). UTF1 was previously described in the context of primed pluripotency and its expression is induced upon transfer of cells to a serum medium [246, 293]. JARID2 plays an important role in PRC2 recruitment and was previously linked with primed pluripotency [294]. Further, MED12 and DPPA2 exhibited similar trends for the ESC Chromatome and POLR2A for the NANOG co-binding protein profiling experiment. We included ASH2L1, CTNNB1, DPPA2, JARID2, MED1, MED12 and UTF1 ChIP-seq experiments as additional chromatin features to predict condition-specific enhancer activity.

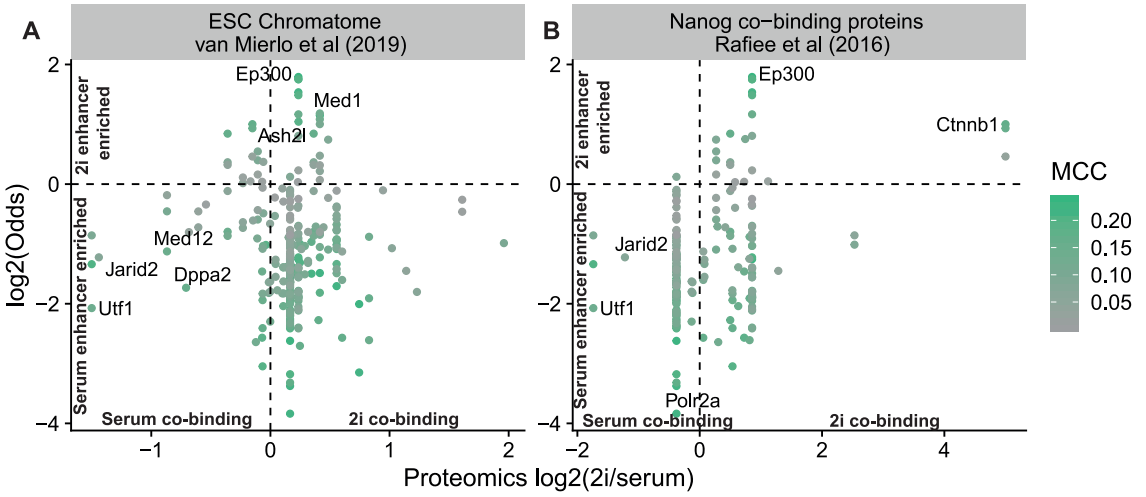


Fig. 3.14: Identification of chromatin associated proteins predictive for differential enhancer activity. Scatterplot showing protein abundance fold-changes between 2i and serum ESC conditions plotted against the differential ChIP-seq peak enrichment in condition-specific enhancer. Proteomics experiments were targeted to **(A)** all chromatin binding proteins (chromatome) and **(B)** NANOG co-binding proteins [204, 205]. The y-axis denotes the log2(Odds) from Fisher's exact test. Colour scale encodes Matthews correlation coefficient (MCC). Selected proteins were highlighted by labelling.

In total, we used 20 different chromatin features, each as a single predictor, to distinguish condition-specific enhancer activities (see Methods). For each predictor, we quantified their predictive power as area under the ROC curve (auROC) and used this as measure of feature importance. This revealed changes in chromatin accessibility as measured by ATAC-seq between 2i and serum ESC as most predictive for condition-specific enhancer activity

(Fig. 3.15). Interestingly, ATAC-seq from samples which were cultured under 2i or serum conditions for longer periods exhibited higher predictive power, than short-term cultures (ATAC-seq LT *auROC* = 0.81, 1 day *auROC* = 0.71, 8h *auROC* = 0.57). Apart from open-chromatin changes, only binding changes of the histone acetylase P300 showed comparably high predictive power (*auROC* = 0.73). This was even higher than for the usual predictors of enhancer activity based either on histone modifications H3K27ac (*auROC* = 0.67) or nascent eRNA transcription (GRO-seq *auROC* = 0.66). Of our previously selected chromatin associated proteins, UTF1 and CTNNB1 showed highest predictive power to distinguish 2i/serum enhancers, which again is in line with the previously published ESC condition-specificity of these proteins [225, 293]. Surprisingly, some features performed barely better than random guessing (*auROC* $\approx$ 0.5) including changes in H3K27me3 or H3K4me1, CpG methylation in 2i ESCs, DPPA2 and MED1 ChIP-seq (Fig. 3.15).

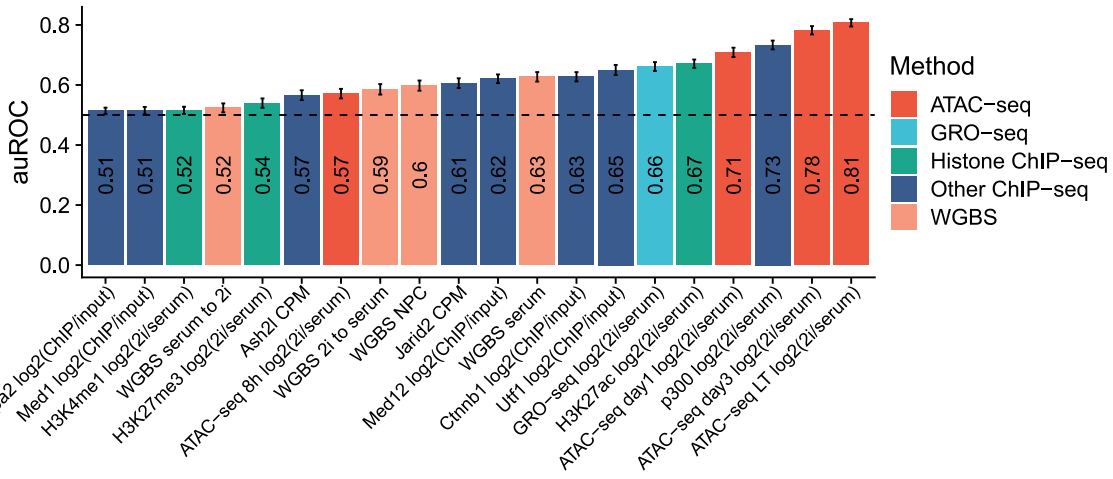


Fig. 3.15: Predictive power of chromatin features for differential enhancer activity. Ranking of 20 different chromatin features according to their predictive power for condition-specific enhancer activity, as measured by area under the ROC curve (auROC). Chromatin features were derived from sequencing methods including ATAC-seq (red), GRO-seq (blue), ChIP-seq profiling of histone modifications (green) or other chromatin associated proteins (purple) and whole-genome bisulfite sequencing (WGBS; orange). X-axis labels include information about exact signal type used for enhancer activity prediction, e.g. $log2(2i/serum)$ for log2 fold-change in signal between conditions or counts per million (CPM). Error bars represents standard deviation which was derived from 100 re-sampling iterations.

In summary, chromatin accessibility changes between 2i and serum ESC conditions were the major predictor for condition-specific enhancer activity. This provides an explanation of how enhancer activity of these sites might be regulated in their endogenous loci.

3.3.11 Prediction of condition-specific enhancer activity using sequence features

Having uncovered chromatin accessibility changes as a major predictor of condition-specific enhancer activity, we investigated what TFs might be involved in this process and how these changes might be encoded in the DNA sequence of the enhancers. To address this, we trained two different machine learning models including GBMs and gkm-SVMs based only on sequence features [211, 208].

GBM sequence feature selection, training and hyperparameter selection

As an initial step to answer these questions, we performed the same feature selection procedure as described in Chapter 2 to determine predictive TF motifs. Briefly first we constrained the TF search space using the two proteomic experiments profiling protein abundance changes between 2i and serum ESC condition [204, 205]. This revealed 137 TFs combining both experiments, 103 TFs from the entire ESC Chromatome and 52 TFs from the Nanog co-binding protein experiment with 18 TFs overlapping between them. We could map 67.2% (92/137) of all TFs to at least one PWM from a recently published consensus PWM dataset [63]. However, this step yielded in total 682 links between TFs and PWMs corresponding to an average of 7.4 TF-PWM links for each TF. To reduce the number of potential TF-PWM links, we scored sequences from condition-specific enhancers, used these scores as classifiers to predict 2i/serum labels and computed the auROC for each PWM as a performance measure. Finally, we pruned all TF-PWM links to keep only the most predictive ones by filtering for the max auROC (Fig. 3.16A).

We found TCF4 ($auROC = 0.697$) and TCF7L1 ($auROC = 0.672$) amongst the most predictive TF motifs for 2i enhancer activity (Fig. 3.16A). Moreover, TCF7L1 showed higher abundance on chromatin as well as NANOG co-binding factor under ESC 2i conditions (Fig. 3.16B). TCF4 was only detectable as Nanog co-binding factor but also showed higher abundance under 2i conditions (Fig. 3.16B). These observations are again in line with previous reports on the importance of WNT signalling for naive (2i) pluripotency [225]. Motifs associated with ARID3A ($auROC = 0.634$) and ESRRB ($auROC = 0.613$) showed high predictive power for 2i enhancers as well as high abundance under 2i condition in both proteomics experiments (Fig. 3.16B). No literature was available about ARID3A under ESC 2i conditions. ESRRB was previously described in context of ESC 2i conditions [199]. In contrast, we found ZIC3 ($auROC = 0.66$), GABPA ($auROC = 0.645$), MAZ ($auROC = 0.644$), ETV1 ($auROC = 0.649$) and CTCF ($auROC = 0.638$) amongst the best predictors for serum

enhancer activity (Fig. 3.16A). However, comparing these factors with our proteomics experiments revealed slightly higher abundances of CTCF, GABPA, MAZ and ETV1 under 2i conditions (Fig. 3.16B). Only ZIC3 showed slightly higher abundance on chromatin under ESC serum conditions. This is in line with previous reports of ZIC3 being involved in the transition from naive to primed pluripotency [295].

Finally, we removed all TF motifs with low predictive power ($auROC < 0.6$) and split the remaining into four distinct feature sets including predictive Chromatome TF features ($n = 17$), Nanog co-binding TF features ($n = 11$), overlapping features ($n = 6$) and the union of all features ($n = 22$). These four feature sets were used to train four separate GBM models. For each of these models we performed a hyperparameter search aiming to optimise the area under the precision recall curve (auPRC) and train a final model (see Methods; Supp.Fig. A.4) [211].

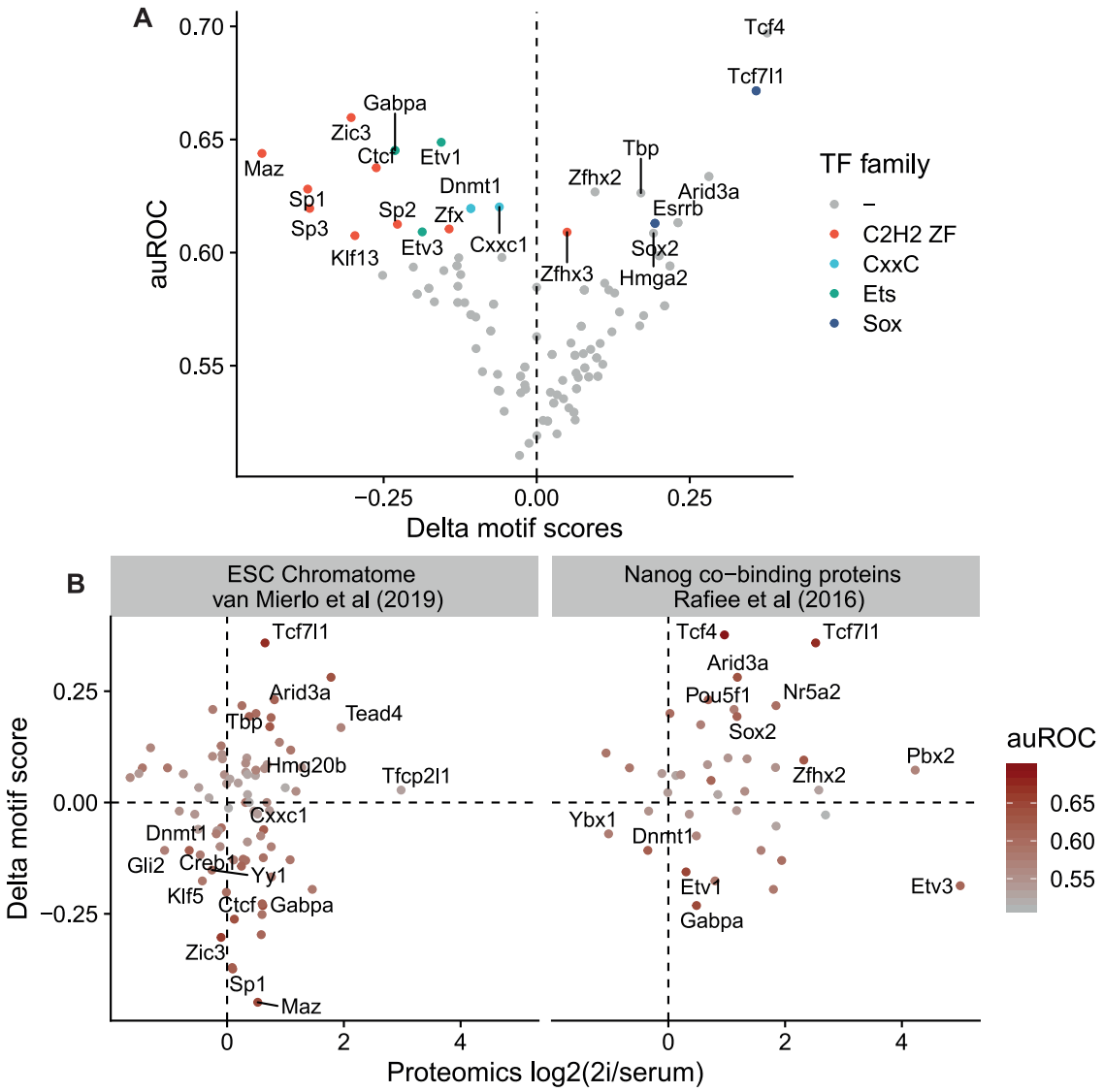


Fig. 3.16: Prioritisation of potential TFs involved in differential enhancer activity.
(A) Volcano-like plot illustrating the difference in median motif scores between 2i- and serum-specific enhancers against predictive power, as measured by auROC, for each TF detected in either proteomics experiment. Only most predictive PWMs for each TFs are shown. TFs with $auROC > 0.58$ were labelled in the plot. Colours indicate the TF family of each predictive PWM ($auROC > 0.6$). **(B)** Scatterplot showing protein abundance fold-changes between 2i and serum ESC conditions plotted against difference in median motif scores. Proteomics experiments were targeted to Chromatome (left panel) and NANOG co-binding proteins (right panel) [204, 205]. Colour scale encodes area under ROC curve (auROC) for the TF motifs.

gkm-SVM training and hyperparameter selection

As in Chapter 2, we initially performed a grid-search to optimise gkm-SVM hyperparameters including k-mer length (L), gap length (K) and number of informative columns (NM; see

Methods) [208]. gkm-SVM models were trained using 10-fold cross-validation and the area under precision-recall curve was recorded for each hyperparameter set. This revealed following parameters set as best one: $L = 8, K = 7, NM = 3$ with an *auPRC* of 0.864. These hyperparameters were used to train a final gkm-SVM model.

GBM and gkm-SVM evaluation on hold-out test set

Next, we evaluated all previously trained models on the same hold-out test set and compared them using five different metrics including area under precision recall curve (auPRC), Matthews correlation coefficient (MCC), balanced accuracy, sensitivity and specificity (see Methods) [194]. Comparing the models using the auPRC as metric revealed that all five models were able to classify 2i enhancers with high precision (Fig. 2.21A). All four GBM models showed comparably high performances (auPRC 0.89) with the exception of the gkm-SVM model which had slightly lower precision (auPRC=0.882). However, this metric is focused on successfully recovering 2i enhancers (true positives) and does not consider the number of successfully classified serum binding sites (true negatives). Therefore, we computed the MCC for each model since this metric takes the entire confusion matrix into account (see Methods). In terms of MCCs, the GBM using all TF motifs as input performed best in accurately distinguishing 2i from serum enhancers (MCC=0.492; Fig. 2.21B). In contrast, the gkm-SVM performed worst using this metric (MCC=0.398) and the GBM using only the Chromatome TFs as input exhibited worst performance among GBM models (MCC=0.453). Nevertheless, we found that GBM models performed comparatively well using this metric (MCC > 0.45). A caveat of MCC as a metric are that it might strongly depend on the applied probability/score threshold which was not directly comparable between GBMs and the gkm-SVM model. To compare our models more comprehensively, we additionally examined the balanced accuracy as another performance measure. We found that the gkm-SVM model showed the highest balanced accuracy (Acc=0.769) and the GBM all TF model the second highest (Acc=0.746; Fig. 2.21C). Overall the GBM trained on the Chromatome TF motifs showed lowest accuracy (Acc=0.724). To better understand MCC and balanced accuracy metrics, we also computed the sensitivity and specificity for all five models. The GBM models showed lower sensitivities but higher specificities than the gkm-SVM model (Fig. 2.21C). Among the GBM models, that trained on all TFs showed highest sensitivity (sens=0.696) as well as specificity (spec=0.796). The GBM model trained on Chromatome TFs showed lowest sensitivity (sens=0.652) and the GBM model trained on NANOG co-binding TFs lowest specificity (spec=0.779). Together, GBM models identified serum enhancers more accurately than 2i enhancers using a fixed probability threshold

($p_{2i} < 0.5$). In contrast, the gkm-SVM model exhibited a more balanced behaviour between sensitivity (sens=0.756) and specificity (spec=0.782). Again, it is important to keep in mind that balanced accuracy, sensitivity and specificity also depend on the chosen probability/score threshold.

In summary, all trained models successfully learned to distinguish 2i and serum enhancers based on sequence features. The GBM model trained using all TF motifs outperformed all other models when assessed by the MCC metric and other GBM models in respect to balanced accuracy/sensitivity/specificity.

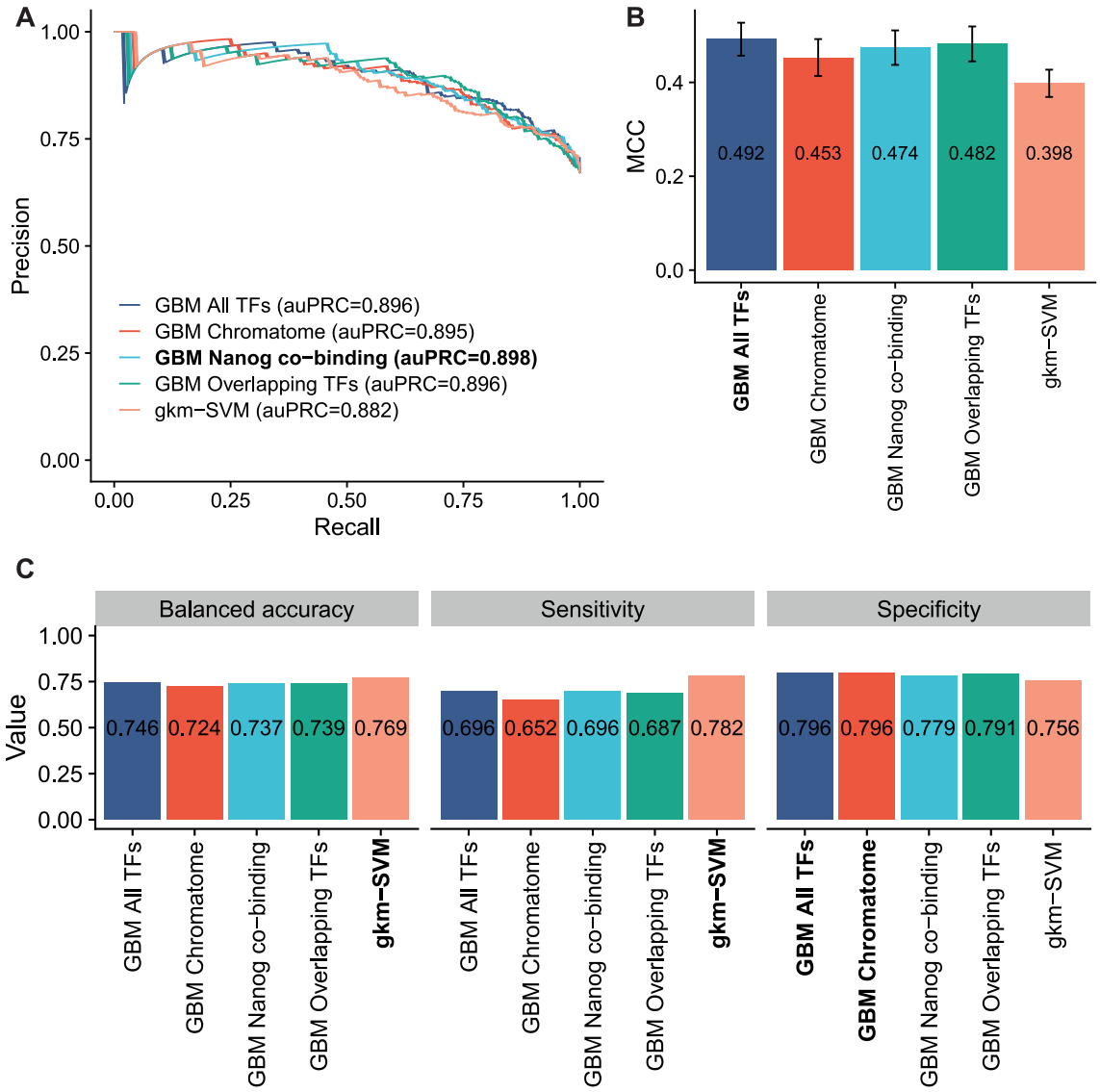


Fig. 3.17: Performance benchmark on hold-out test set of classification models using different metrics. **(A)** Precision-recall curves for all trained models including GBM all TFs (purple), GBM Chromatome (red), GBM Nanog co-binding (light blue), GBM Overlapping TFs (green) and gkm-SVM (orange). Precision-recall curves were computed on a hold-out test set. Area under precision recall curve (auPRC) for each model is shown in the legend. **(B)** Average Matthews correlation coefficient (MCC) for all trained models evaluated on a hold-out test set. Error bars represent standard deviation of MCCs calculated from 100 re-sampling iterations. Shown MCC values were rounded to three digits after decimal point. **(C)** Balanced accuracy, sensitivity (true positive rate) and specificity (true negative rate) for all trained models evaluated on a hold-out test set. Shown performance metrics values were rounded to three digits after decimal point.

GBM feature importance and validation of important features using ChIP-seq

Having established that the DNA sequences underlying enhancers contain enough information to distinguish 2i from serum enhancer elements, we investigated which sequence features (e.g. TF motifs or dinucleotides) might be the most important. To answer this question, we made use of the best performing GBM model and computed an importance score for each model feature (see Methods). This revealed the motif for TCF4 as the most important model feature by far (Fig. 3.18A). Moreover, we found PWMs associated with ZIC3, TCF7L1, ESRRB and CTCF among the top five most important TF motif features, consistent with the analysis of single PWMs (Fig. 3.16A).

From all dinucleotide features, we identified AA and TT as most important (Fig. 3.18A). Closer examination of AA/TT dinucleotide features revealed higher median frequencies in 2i enhancers compared with serum enhancers ($AA_{2i} = 8.5\%$, $AA_{serum} = 7\%$, $TT_{2i} = 8.2\%$, $TT_{serum} = 6.8\%$). These dinucleotides might account either for missing TF motifs, simple repetitive sequences or higher order sequence features.

Next, we examined the relationship between STARR-seq signal and motif scores to better understand factors contributing to differential activity (Fig. 3.18B). This revealed that ESRRB, TCF4 and TCF7L1 showed higher motif scores for 2i enhancers and were therefore more specific to these enhancers. TCF4 motif scores increased with increasing 2i enhancer activity and decreased with increasing serum enhancer activity. TCF7L1 showed a similar relationship and ESRRB exhibited a more complex pattern. In contrast, ZIC3 showed constantly higher motif scores for serum enhancers independently of the STARR-seq signal (Fig. 3.18B). These observations were in line with our previous proteomics results above for which ESRRB, TCF4, TCF7L1 showed higher abundances under 2i conditions and ZIC3 higher abundance under serum conditions (Fig. 3.16B). Finding the WNT signalling associated TFs, TCF4 and TCF7L1, among the most important features for 2i enhancers is in line with the literature emphasising the importance of WNT signalling for naive pluripotency (2i) [225]. Moreover, ESRRB was previously described as important TF involved in naive pluripotency [296, 199]. On the other hand, ZIC3 being important for serum enhancer prediction is in line with previous reports of its involvement in primed pluripotency [297, 295].

Together, we identified TCF4, TCF7L1 and ESRRB as most important predictors of 2i enhancer activity and ZIC3 as the most important for serum-specific enhancer activity.

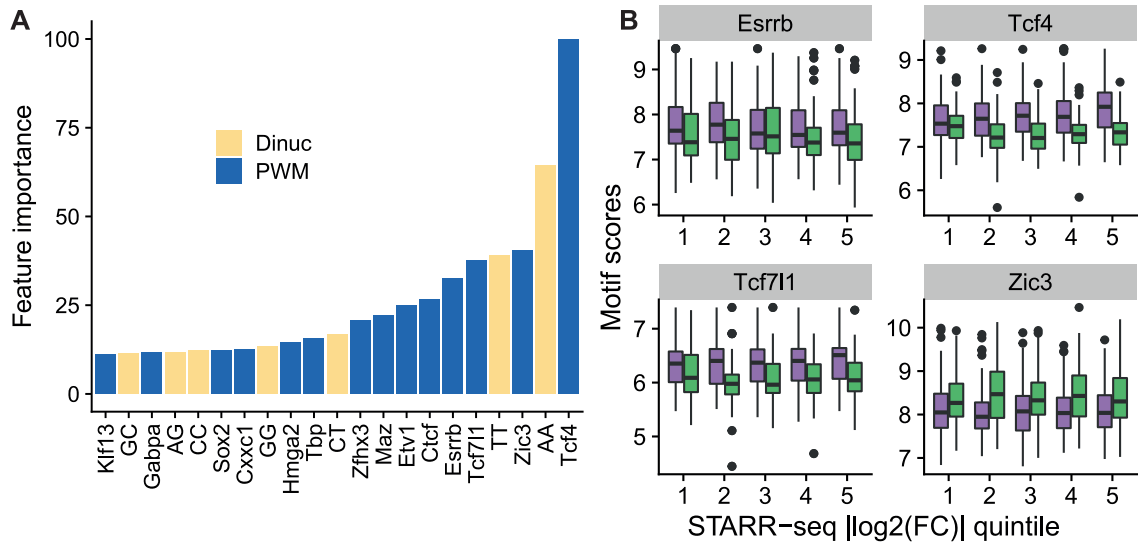


Fig. 3.18: Feature importance ranking for GBM trained on all proteomics TFs. **(A)** Feature importance ranking showing the 20 most important features. Colour indicates feature type, PWM (blue) or dinucleotide frequency (yellow). **(B)** Boxplot showing motif score distributions for the top four most important PWM features: TCF4, ZIC3, TCF7L1, and ESRRB. Motif scores were stratified into quintiles according to the absolute STARR-seq signal ($|log2(FC)|$).

Having identified important predictive TF motifs, we sought to validate binding of the identified TFs to our condition-specific enhancer sets using publicly available ChIP-seq samples. We were unable to find any publicly available ChIP-seq data for TCF4 in the two major sequencing data repositories (Gene Expression Omnibus and ArrayExpress). However, we found a publicly available ChIP-seq dataset for TCF7L1, the other important TF associated with WNT signalling [298]. Analysis of this dataset revealed TCF7L1 binding to both enhancer classes but higher average binding to 2i-specific enhancers (Fig. 3.19A). More generally, 2i-specific enhancers exhibited significantly higher levels of TCF7L1 binding compared with serum enhancers ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 3.19B). This provides additional evidence for TCF7L1 being an important regulator for differential enhancer activity.

We also found publicly available ChIP-seq data for ZIC3, the second most important TF feature which was predicted to be more specific for serum enhancers [299]. The ZIC3 ChIP-seq exhibited signal enrichment at both enhancer types but as expected from our previous analysis higher average binding at serum-specific enhancers (Fig. 3.19C). Comparison of signal enrichment within condition-specific enhancers confirmed significant higher binding of ZIC3 to serum enhancers ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test; Fig. 3.19D). These

observations point towards ZIC3 being an important regulator for serum-specific enhancer activity.

Finally, we also investigated differential binding of ESRRB to our condition-specific enhancers and therefore used the same ESRRB ChIP-seq from 2i and serum cultured ESCs as in Chapter 2 [199]. Again, this confirmed our expectation of higher average ESRRB binding to 2i-specific enhancers than serum enhancers under ESC 2i conditions (Fig. 3.19E). However, ChIP-seq profiling of ESRRB under ESC serum condition showed dramatic reduction in average signal enrichment at both enhancer types. Quantification of signal changes between ESRRB ChIP-seq from 2i and serum ESCs confirmed binding of ESRRB to both enhancer classes was specific to 2i conditions (Fig. 3.19F). Moreover, this effect was significantly higher at 2i enhancers than serum enhancers ($p < 2.22 \cdot 10^{-16}$, one-sided Wilcoxon test). Hence, this suggests that ESRRB is involved in the regulation of differential enhancer activity between 2i and serum conditions. This is in line with our previous observation that ESRRB might be the most important regulator of enhancer activity in ESC 2i conditions (see Chapter 2 for extensive discussion).

In summary, here we validated the predicted binding of TCF7L1, ESRRB and ZIC3 to condition-specific enhancers, thus providing evidence for them as potential regulators of condition-specific enhancer activity. However, more experiments are necessary to validate their putative role as regulators of *bona fide* enhancer activity.

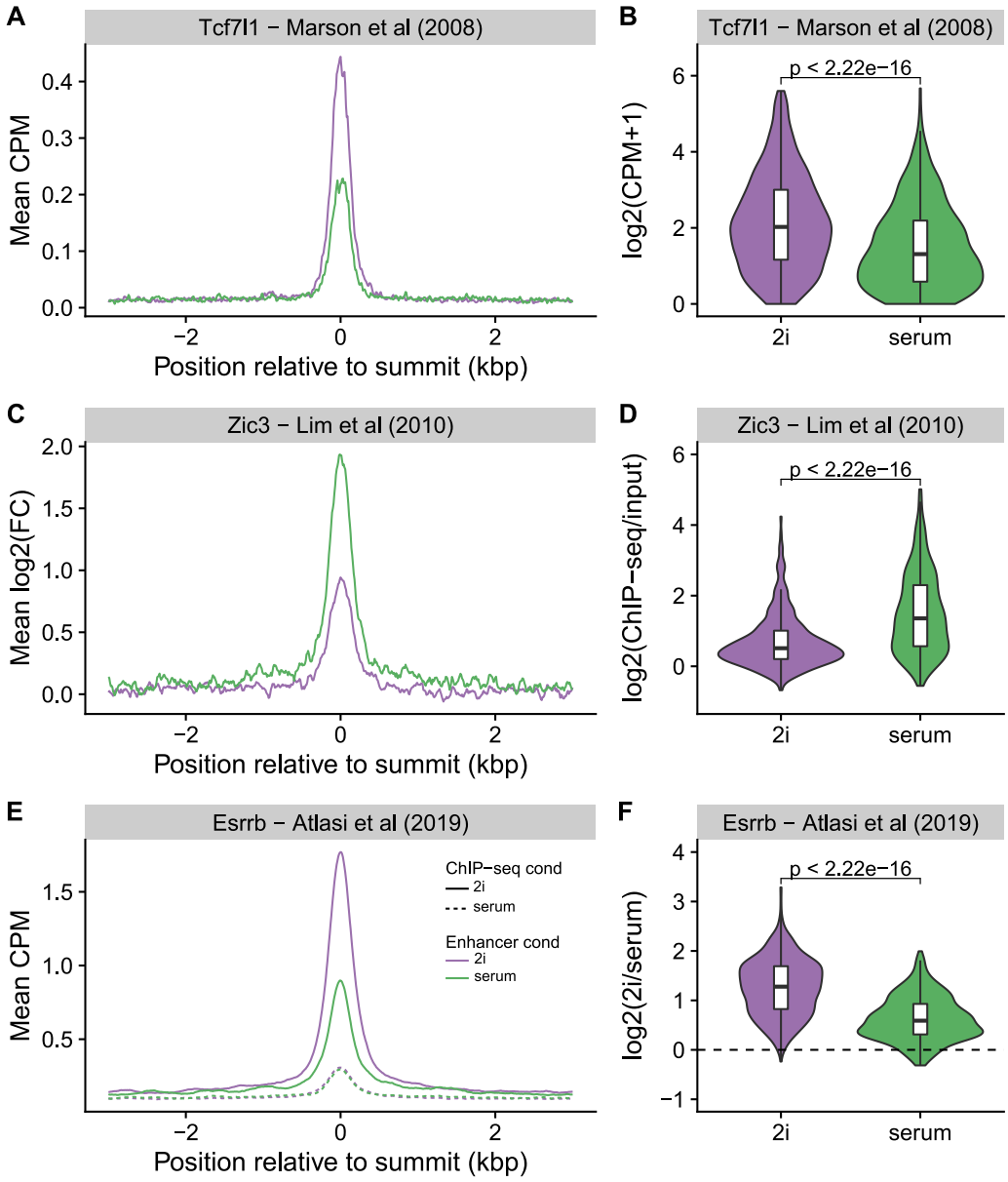


Fig. 3.19: TCF7L1, ZIC3 and ESRRB binding, as measured by ChIP-seq, at 2i- and serum-specific enhancers. **(A)** Metaprofile showing average TCF7L1 ChIP-seq signal, as CPM, at ±3kbps around 2i (purple) and serum (green) enhancers [298]. **(B)** Violin plot comparison of TCF7L1 ChIP-seq signal (CPM) within 2i (purple) and serum (green) enhancers. P-value from one-sided Wilcoxon test. **(C)** Metaprofile showing average ZIC3 ChIP-seq signal, as log2 fold-change over input, at 2i and serum enhancers [299]. **(D)** Violin plot comparison of ZIC3 ChIP-seq signal (log2 fold-change over input) within 2i and serum enhancers. P-value from one-sided Wilcoxon test. **(E)** Metaprofile showing average ESRRB ChIP-seq signal, as CPM, at 2i and serum enhancers [199]. Condition-specific ESRRB ChIP-seq samples for 2i (solid line) and serum (dotted line) mESCs are shown. **(F)** Violin plot comparison of ESRRB binding changes between ChIP-seq from 2i and serum mESC, within 2i and serum enhancers. P-value from one-sided Wilcoxon test.

3.4 Discussion

In this chapter, we investigated the features underlying differential enhancer activities exhibited by NANOG binding sites in naive (2i) and primed (serum) mouse ESCs. To this end, we used a recently published protocol combining ChIP with the high-throughput enhancer assay STARR-seq to identify *bona fide* transcriptional enhancers [178]. Application of this technique in 2i and serum ESCs followed by differential enhancer activity analysis revealed enhancers displaying condition-specific activity. We comprehensively characterised these condition-specific enhancers in terms of TF binding, chromatin landscape, DNA methylation, nascent eRNA transcription, chromatin-chromatin interactions and target gene expression. Finally, we used two machine learning algorithms to uncover sequence features (e.g. TF motifs) predictive for the observed differential enhancer activity and validated predicted regulators with publicly available ChIP-seq data.

First, we defined high-confidence NANOG binding sites which displayed fragment coverage in our STARR-seq vector library (see Chapter 2 Section 2.3.1). To achieve this, we combined NANOG binding sites from three different ESC conditions including naive (2i), primed (serum) and epiblast stem cells. However, the coverage filtering was based on a STARR-seq vector library prepared from a NANOG ChIP profiled under 2i conditions. Therefore, we examined whether NANOG binding sites might identify only those bound in 2i. Previous studies reported dynamic binding of NANOG in different pluripotency conditions as well as during differentiation [216, 222]. For example, Galonska *et al* compared NANOG binding in serum with 2i conditions and found 14,796 differential sites or De Kumar *et al* revealing dynamic NANOG binding upon induced differentiation with RA [216, 222]. Hence, creating a STARR-seq vector library for all three conditions (2i/serum/EpiSC) and then cross-transfecting them, would broaden the search space for putative enhancer elements and allow for a more comprehensive investigation of condition-specific enhancer activity. In our approach, we only examined differential enhancer activity of NANOG bound sites specific to 2i or shared between conditions. Thus, NANOG binding sites which were specifically bound in serum or EpiSC were not included in our analysis.
Next, we defined *bona fide* transcriptional enhancer specific to the 2i or serum condition by performing differential count analysis using edgeR (see Methods) [195]. This uncovered 768 Nanog binding sites with specific enhancer activities in 2i and 481 in serum conditions. As discussed in Chapter 2, enhancer calling from our data sets might be negatively influenced by the use of an older version of the STARR-seq screening vector and the lack of UMIs in our

experiment [176, 219, 184]. Nevertheless, we were able to detect significant differentially active enhancers between the two conditions which might explain at least some of the gene expression differences.

To ensure the validity of our enhancers, we compared our condition-specific enhancers with previously published transcriptional enhancers, which were validated using a dual-luciferase reporter assay. Indeed, two of our 2i-specific enhancers could be matched with two previously described regions associated with *Prdm14* and *Jarid2* regulation (Fig. 3.3 & Data not shown) [216, 199]. We could not recover known serum enhancers, such as the *Oct4* proximal enhancer and the *Fgf5* associated enhancer (Data not shown) [216, 300]. However, manual inspection of our STARR-seq data clearly confirmed condition-specific enhancer activity changes at our serum-specific enhancers (Fig. 3.4). The most likely explanation for these missing enhancers is that they correspond to serum-specific NANOG binding sites and were thus not represented in our vector library. This highlights the importance of generating and assaying condition-specific vector libraries to obtain a more comprehensive picture of differential enhancer activity.

We compared various published regulatory genomic datasets, including ChIP-seq, ATAC-seq, WGBS, GRO-seq, Capture-HiC, with our condition-specific enhancer sets in order to collect additional evidence supporting their differential activity. Descriptive analyses of all these datasets confirmed condition-specific differences between enhancer classes including accessibility, histone modifications, nascent eRNA transcription, methylation and TF engagement. Thus, integration of publicly available data with our condition-specific enhancers supported their differential activity and all our observations were in line with expectations from previous publications [161, 42]. Beyond descriptive analysis, we aimed to prioritise these chromatin features by their predictive power for condition-specific enhancer activity. This revealed that chromatin accessibility changes as measured by ATAC-seq was is by far the best predictor ($auROC = 0.81$). Changes in the epigenetic writer P300 and its associated histone mark H3K27ac showed lower predictive power (P300: $auROC = 0.73$, H3K27ac: $auROC = 0.67$). These observations were supported by previous reports that condition-specific accessible and H3K27ac marked regions showed *bona fide* enhancer activity in a dual-luciferase reporter assay alone [199]. In contrast, regions only showing dynamic accessibility changes showed basal levels of transcription [199]. In addition to the chromatin features discussed above, we investigated their ability to predict condition-specific enhancer activity. Initially we complemented the set of chromatin features with potentially predictive chromatin associated proteins

by integrating publicly available chromatin profiling proteomic data with ChIP-seq data [204, 205]. This analysis uncovered 4 chromatin associated proteins, CTNNB1 ($auROC = 0.63$), JARID2 ($auROC = 0.61$), MED12 ($auROC = 0.62$) and UTF1 ($auROC = 0.65$) with the ability to distinguish condition-specific enhancers. These factors were previously described in the literature as being involved in the regulation of naive/primed pluripotency. For example, CTNNB1 was shown to be crucial under 2i conditions to overcome the repressive effect of TCF3 and promote pluripotency [301]. In contrast, JARID2, a subunit of the PRC2 complex, was shown to be involved in primed pluripotency, and ESCs lacking JARID2 failed to undergo differentiation [294]. UTF1 was shown to be highly up-regulated upon transfer of ESCs from 2i to serum and was previously linked to primed pluripotency via the PRC2 complex [246, 293].
As in Chapter 2, we performed predictive modelling of condition-specific enhancer activity using sequence features only. We trained GBM models using proteomics data combined with PWMs and also trained a gkm-SVM model directly from enhancer sequences. Here, we were unable to train a CNN model due to the small dataset size. Consequently, we could not compute a base-pair level importance score using DeepLIFT and derive novel sequence features as in Chapter 2 [212]. Nevertheless, GBM and gkm-SVM models showed good performance in distinguishing 2i from serum enhancers just using sequence features. Moreover, a recent publication described "GkmExplain" which implements a way to derive base-pair level importance scores from a gkm-SVM model [302]. We aim to apply this tool in future.

Predictive modelling of condition specificity using GBMs, uncovered TCF3/TCF7L1 and TCF4/TCF7L2, two TFs involved WNT signalling, as important predictors of 2i enhancer activity. For TCF4, we did not find any supporting literature describing its role in pluripotency or ESC biology; however, we showed that TCF4 is expressed at low levels in ESC cells and was identified as Nanog co-binding TF using ChIP-SICAP [204]. On the other hand, TCF3 was previously described as integral part of the core-regulatory circuitry linking pluripotency to differentiation [303]. Its impact on gene expression depends on the co-localisation with CTNNB1: bound together, TCF3 functions as activator but bound alone, it is a repressor. Hence, inhibition of GSK-3 under 2i culture condition induces activation of canonical WNT signalling, resulting in co-localisation of CTNNB1 and TCF3 which ultimately promotes naive pluripotency [304]. Changing culture condition from 2i to serum would result in the lack of canonical WNT signalling and therefore might lead to repression of 2i-specific enhancers by TCF3. Moreover, this analysis uncovered ESRRB as important predictor of

2i-specific enhancer activity. This is consistent with previously published research and our work in Chapter 2 [199]. Finally, we found the ZIC3 motif to be the most predictive of enhancer activity in serum conditions. Previous research showed that ZIC3 is required for maintaining pluripotency and is a direct regulator of *Nanog* expression [297, 299]. Moreover, ZIC3 has been indicated as a reprogramming factor together with ESRRB for inducing pluripotency [251]. In line with our observations, ZIC3 was implicated as regulating the transition from naive to primed pluripotency [295].

In summary, in this chapter we uncovered *bona fide* transcriptional enhancers exhibiting differential activity between naive (2i) and primed (serum) pluripotency. These elements were comprehensibly characterised revealing changes in chromatin accessibility, H3K27ac and P300 between conditions as major predictors for differential enhancer activity. Predictive modelling using sequence features uncovered the two WNT related TFs TCF3/TCF7L1 and TCF4 as well as ESRRB as major predictors for 2i-specific activity and ZIC3 as a major determinant of serum activity. As in Chapter 2, additional wet-lab experiments are necessary to validate these predictions and reveal potential mechanisms involved in regulating differential enhancer activity. Our proposed workflow of combining differential ChIP STARR-seq with machine learning approaches will help to uncover how condition specific enhancer activity is encoded in DNA sequence and help to prioritise candidate TFs as master regulators of compared conditions.

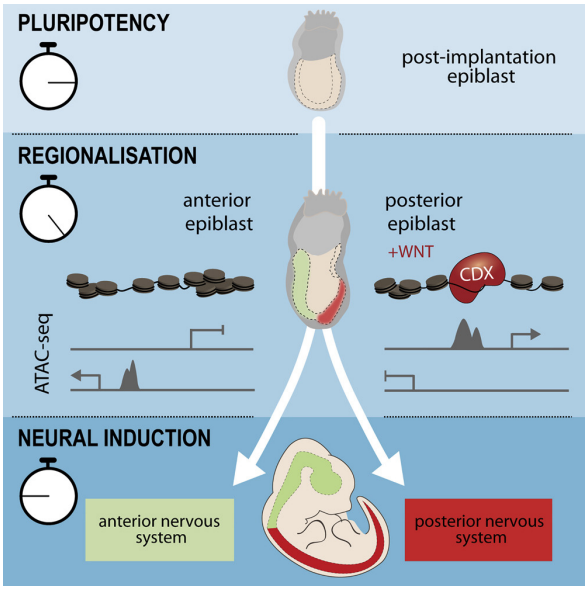
4

Chromatin accessibility dynamics during early neurogenesis reveal timing of axis allocation and neural induction

Neural induction in vertebrates generates a central nervous system that extends the anterior-posterior length of the body. The prevailing view is that neural induction initially generates cells with anterior (forebrain) identity, caudalising signals then converting a proportion of these to posterior fates (spinal cord). Here we challenged this model using genome-wide open-chromatin profiling by ATAC-seq in both an *in vitro* model system of this process as well as *in vivo*. Contrary to the established model, these data indicated that for spinal cord induction, cells commit to a posterior regional identity before acquiring neural identity. Additionally, we identified distinct sets of regulatory elements required for the regional identity of neural cell identities. In combination with extensive mining of published sequencing datasets, including RNA-seq, ChIP-seq and ATAC-seq, we uncovered key transcription factors and their activating/repressing contributions to this process. This highlighted the role of CDX2 in specifying posterior identity for spinal cord development. Perturbation experiments supported these data and confirmed the role of key transcription factors involved in inducing regional identity. In summary, we showed that cells initially acquire regional identity and then neural identity prompting a revision of the current model. This is further supported by the proposed dual evolutionary origin of vertebrate central nervous system. ■

Declaration The work presented in this chapter was a collaborative work between the Briscoe and Luscombe labs at the Francis Crick Institute.
Vicki Metzis and James Briscoe conceived the project. V.M. performed wet lab experiments including ESC differentiation, preparation of ATAC-seq and ChIP-seq libraries and qRT-PCR experiments. V.M. analysed the qRT-PCR shown and prepared all schematics used in this chapter. I (= Sebastian Steinhauser) analysed all high-throughput sequencing data including mRNA-seq, ATAC-seq and ChIP-seq as well as additional data sources presented in this chapter. Figures relating to this data were produced with code I implemented. Other contributions included: Thomas Watson performed flow cytometry experiments and Teresa Rayon performed confocal imaging.

The presented work was published as:
Vicki Metzis*, Sebastian Steinhauser*, Edvinas Pakanavicius, Mina Gouti, Despina Stamataki, Kenzo Ivanovitch, Thomas Watson, Teresa Rayon, S. Neda Mousavy Gharavy, Robin Lovell-Badge, Nicholas M. Luscombe, and James Briscoe.
Nervous System Regionalization Entails Axial Allocation before Neural Differentiation. *Cell*, 175(4):1105–1118, 2018. (* = Authors contributed equally)



Graphical abstract was taken from [305].

4.1 Introduction

Gene expression programs need to be spatiotemporally regulated in order to create a viable organism from a fertilised egg during development. It is well established that distal regulatory regions, so called enhancer elements, encode this information in form of transcription factor (TF) binding sites. Here, we focus on the question of how these elements regulate gene expression during early neural development in vertebrates and what are TFs involved in this. The development of the nervous system, a process referred to as neural induction, begins at gastrulation, the process which determines the formation of the three primary germ layers of the embryo [306]. Neural induction continues throughout embryogenesis, producing a nervous system that forms along the rostro-caudal or anterior posterior axis (AP-axis). During this process, cells differentiate towards a neural cell fate producing neural progenitor cells (NPCs) [306]. These NPCs give rise to various cell types depending on their location in the rostral-caudal regions of the developing nervous system.

Different hypotheses have been proposed to describe how cells acquire neural as well as regional identity. In 1924, Otto Mangold suggested that distinct mechanisms are responsible to generate distinct regions of the nervous systems [307]. However, the prevailing view is summarised in the Activation-Transformation hypothesis [308] (Fig. 4.1 *upper panel*). According to this model, during "Activation" describes that cells initially differentiate towards an anterior neural identity (forebrain). Subsequently during "Transformation", these anterior neural cells give rise to a plethora of other neural cells comprising the midbrain, hindbrain and spinal cord ("Transformation"). In this model all neural cells along the anterior-posterior nervous system would originate from a forebrain neural identity [308].

In contrast to the above model however, the posterior spinal cord (thoracic level and lower) and anterior nervous system (fore-, mid-, hindbrain and anterior spinal cord) develop from anatomical and molecular distinct cell types [309]. These observations were summarised in an alternative hypothesis called the "Dual origin of neural progenitors" hypothesis (Fig. 4.1 *lower panel*). Here, cells from the anterior epiblast will give rise to the anterior nervous system, whereas, the posterior spinal cord originates from a bipotent stem cell population within the posterior epiblast, the neuromesodermal progenitors (NMPs) [310, 311]. These NMPs can differentiate into the neuroectoderm as well as the trunk mesoderm and were identified as the driving force of productive axis elongation. In contrast to the anterior epiblast, cells in the posterior part are exposed to WNT and Fibroblast growth factor (FGF) signalling

which appear to induce the NMP state. Hallmarks of this state are the expression of TFs such as SRY-Box Transcription Factor 2 (SOX2), T/Brachyury (T/BRA) and Caudal Type Homeobox (CDX)1, 2, 4 [310, 311]. The importance of this cell population is highlighted by impaired axis elongation in mouse embryos upon the knock out (KO) of *T/Bra*, *Cdx* genes or the absence of WNT signalling [312, 313, 314, 315].

Recent advances in *in vitro* model systems can recreate these complex developmental processes by applying timely pulses of the required signalling pathways to direct embryonic stem cell (ESC) differentiation [310]. ESCs are removed from pluripotent conditions and grown in a basal medium exposed to FGF signalling for at least 48h to differentiate them into epiblast like stem cells (EpiSC). These EpiSC can be used to differentiate cells towards an anterior neural identity in the presence of neuralising signals. Exposure to retinoic acid (RA) allows these cells to adopt a hindbrain NPC identity while the addition of sonic hedgehog (SHH) signalling promotes ventral neural fates. In the absence of added RA, EpiSC adopt a fore-/midbrain NPC identity. In contrast, to generate the more posterior parts of the nervous system, such as the spinal cord, EpiSC must first be caudalised by exposure to WNT and FGF signalling for 24h. This inductive event induces a population of cells that resemble NMPs that form in the tailbud in vivo [310, 316, 317]. These *in vitro* NMPs can be further differentiated into posterior spinal cord NPCs by addition of RA and SHH [310]. Resulting *in vitro* spinal cord NPCs will be marked by expression of posterior *Homeobox (Hox)* genes. Single-cell RNA-seq (scRNA-seq) studies of such *in vitro* generated NPCs and corresponding *in vivo* populations have confirmed their concordance [311, 318]. The observation that distinct NPC cell types can be differentiated *in vitro* from ESCs have challenged the "Activation-Transformation" hypothesis and supported Mangold's model, in which distinct mechanisms lead to different nervous system regions [307].

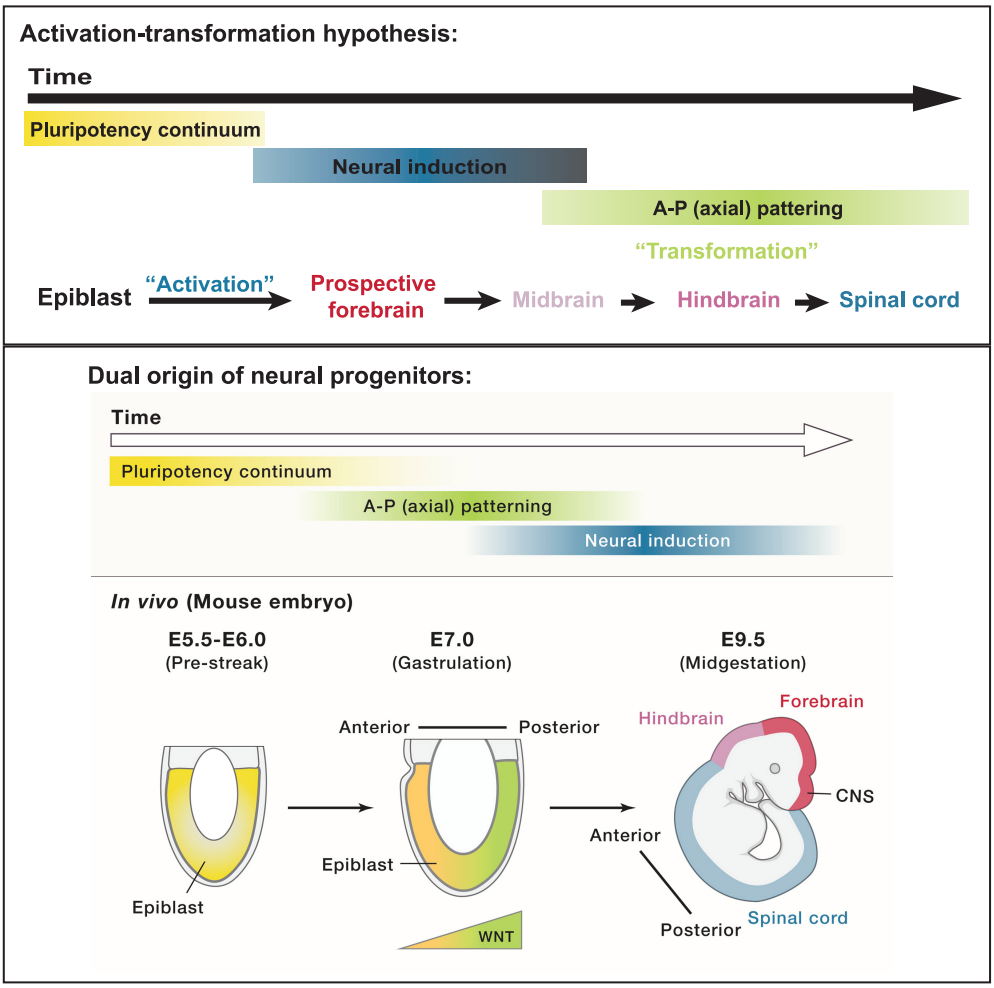


Fig. 4.1: Proposed hypothesis for neural induction and AP-axis pattering during early neurogenesis. *Upper panel:* Schematic diagram summarising the Activation-Transformation hypothesis. According to this hypothesis, cells are initially exposed to signals inducing a forebrain-like neural identity ("Activation"). These cells then acquire other regional identities along the anterior-posterior (A-P) axis by exposure to different signals ("Transformation"). Figure concept was adapted from [308]. *Lower panel:* Schematic diagram summarising the Dual-origin of neural progenitors hypothesis. Pluripotent epiblast cells are exposed to a WNT signalling gradient along the AP-axis during gastrulation (E7.0). The WNT gradient differentiates cells in the posterior epiblast to neuromesodermal progenitor cells (NMPs) which are subsequently committed to a regional posterior identity. These NMPs then give rise to the neuroectoderm, as well as the mesoderm in the developing tail/posterior spinal cord. In contrast, anterior epiblast cells can give rise to neuroectoderm of the central nervous system. Figure was adapted from [319].

In the work presented here, we took advantage of the directed differentiation of ESCs into defined NPCs *in vitro* model to understand the sequence of events, necessary to develop NPCs with different regional identities. Classic embryology experiments are strongly dependent on observable morphological changes and ignore molecular changes such as gene expression

and chromatin dynamics in the embryo [312, 313]. Although it is now becoming possible to apply sequencing techniques to cells isolated from dissected embryos, such approaches often lack the necessary cellular and temporal resolution. Furthermore, perturbation experiments are often either time-consuming or impossible using *in vivo* models. In contrast, the *in vitro* model offers a well-defined and controlled experimental system to direct cells towards different cell fates through the application of well-defined signals. Perturbation are possible by delaying the activation of one signalling pathway over another signal. Previously published research focused on gene expression measurements as a better proxy to define cellular identities [310, 311, 318]. However, recent studies show that measuring the accessibility of distal regulatory regions, also known as enhancers, by profiling for open chromatin profiling serve as a more accurate definitions of cell identity than bulk gene expression profiling [167, 168, 169]. The reasoning behind this is a well-established concept of developmental biology, that important developmental genes (e.g. *Hox* genes) need to be tightly spatiotemporally regulated, achieved through the use of specific enhancers [42, 320]. Within the neural tube, several key transcription factors important in neural development are expressed throughout the nervous system, yet under the control of multiple regulatory elements. One example highlighting the advantage of enhancer accessibility measurements over gene expression is *Sox2*, which is broadly expressed along the AP axis but shows differential regulation by distinct groups of cell-type specific enhancers[321, 322, 323, 324]. Hence, enhancer accessibility needs to be cell-type specific and therefore predictive for cellular identity.

In this chapter, we asked how cells adopt a neural identity, and how cells adopt different regional identities in the anterior versus posterior part of the nervous system. To assess this, we used ATAC-seq, a genome-scale method to profile chromatin accessibility dynamics during *in vitro* differentiation of NPCs with different regional identities along the AP-axis [131]. This revealed sets of enhancer elements specific to different regional identities. To confirm their importance during neural development *in vivo*, we measured their accessibility using ATAC-seq on FACS sorted cells from mouse embryos. In combination with extensive mining of published sequencing data, including RNA-seq, ChIP-seq and ATAC-seq, we use the temporal resolution of the in vitro differentiation to define when cells adopt different AP fates. This revealed that in contrast to the textbook view, cells adopt either anterior or posterior identities before they commit to a neural identity. Moreover, we uncovered key TFs and their regulatory contributions to this process. In doing so, we highlighted the importance of the TF CDX2 for providing cells with the competence to make spinal cord.

4.2 Methods

4.2.1 ATAC-seq and ChIP-seq

All ATAC-seq experiments were performed by Vicki Metzis (Postdoc in the Briscoe Lab) and sequenced by the Advanced Sequencing Platform at the Francis Crick Institute. SOX2 ChIP-seq for day five hindbrain and spinal cord NPCs was also performed by Vicki Metzis and sequenced at the Crick. For exact description of experimental procedures see [305].

4.2.2 ATAC-seq analysis

Preprocessing

For each ATAC-seq sample, sequencing adapters and poor quality bases were trimmed from sequencing reads using trim_galore with default settings (*https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*). Reads were mapped to the *mm10* reference genome using bowtie2 (parameters -X 2000 –sensitive-local) [122]. Unmapped reads, reads with low quality mapping scores ($MAPQ < 30$), reads mapping to chrM, and unpaired reads were removed from further analysis. PCR duplicates were removed using Piccard. For analyses measuring insertion level (e.g., TF footprinting) we shifted plus-strand insertions by +4bp and minus-strand insertions by -5bp [131]. Publicly available *in vivo* ATAC-seq samples were downloaded from GEO and listed in Supp. Table B.12.

Peak calling and annotation

Coverage tracks were computed as fragments per million per base pair (FPM) using deepTools bamCoverage (parameters –scaleFactor 10^6/Library size -bs 1 -extendReads -samFlagInclude 66 -ignoreDuplicates) [124] and enriched genomic loci were identified using MACS2 (parameters -g mm -p 0.01 -nomodel -f BAMPE) [185] For each condition, we combined biological replicates by computing the irreproducible discovery rate (IDR) and using a threshold of $IDR \leq 0.1$ [188]. We defined a consensus peak set across all conditions by merging overlapping peaks across different conditions using BEDTools merge [189]. The coverage of the consensus peaks was calculated for each condition using featureCounts (parameters -F SAF -p -ignoreDup) [187], with higher coverage indicating greater chromatin accessibility.
The consensus peaks were assigned to nearby genes using ChIPseeker annotatePeak [196] with GENCODE release M14 [197] and GREAT (default settings) [128]. GREAT assigns regulatory regions, as proximal, to a gene by overlapping it with a region spanning 5kb

upstream and 1kb downstream of the genes TSS (irrespective of other genes). Additionally, distal regulatory regions were assigned to genes within a maximum of 1,000kb [128].

Differential peak calling

To identify differentially accessible ATAC-seq peaks between conditions, we used the above count table for the consensus peak set as input data for DESeq2 (default settings); pairwise comparisons were made between ESC day 0 (D0) and each WT *in vitro* confdition [156]. Statistically significant differential peaks were identified as those with $log2(FC) > 1$ and $FDR < 0.01$.

Self-organising map clustering

Differentially accessible ATAC-seq peaks were clustered using a self-organising map (SOM). For this, the DESeq2 normalised count data for each differentially accessible peak was transformed into z-scores ($z = \frac{(x - mean(x))}{sd(x)}$). The resulting z-score matrix was used as input for the SOM with a 5x5 cell grid, hexagonal topology and Gaussian neighbourhood function [29]. We found that 5x5 provides the best separation of differential ATAC-seq peaks, though different grid sizes produced similar results. We classified SOM clusters into the conditions in which the differential peaks were most accessible. Resulting in 10 distinct clusters which included: Hindbrain, Neural, Anterior, Hindbrain/Spinal cord, Spinal cord, NMP/Spinal cord, NMP, EpiSC, Unknown, Anterior/Hindbrain and Anterior.

Comparison between *in vivo* vs *in vitro* ATAC-seq

To compare our D5A and D5SC *in vitro* neural progenitors to *in vivo*-derived progenitors, we mapped the ATAC-seq fragment coverage obtained from the *in vivo* datasets to the consensus peak set using featureCounts with the same parameters as above. We normalised the counts from the two datasets using DESeq2 with default settings. Next, we selected the consensus peaks belonging to the anterior, neural and spinal cord SOM clusters, and for each condition-pair, we compared the log2 fold-change distribution using a two-sided Wilcoxon rank sum test. To plot the ATAC-seq meta profiles we included consensus peaks enriched in both *in vitro* and the corresponding *in vivo* sample ($abs(log2(FC)) > 0.5$).
To compare NMP, NMP/SC and Epi SOM cluster peak sets to the *in vivo* datasets, MACS2 called peaks from E6.0, E7.2, E7.5P [325] and E9.5SC and E9.5A neural progenitors were compared to each SOM cluster by computing the peak overlap using findOverlaps from

the GenomicRanges package [191]. The resulting proportion of regions overlapping was calculated relative to the size of the SOM cluster.

Comparison with ENCODE and VISTA enhancer database

DNase hypersensitive sites (DHS) in the *mm10* reference genome were obtained from the ENCODE data portal *(https://www.encodeproject.org/)* (Supp. Table B.16) [326, 327]. DHS regions were overlapped with our consensus peak set using GenomicRanges findOverlaps. VISTA enhancer regions were downloaded from the ENCODE data portal (Supp. Table B.16). Enhancers were overlapped with our consensus peak set and enhancer enrichments were assessed with a one-sided binomial test [128]. P-values were corrected for multiple testing using the Benjamini-Hochberg procedure [328].

TF motif enrichment analysis

TF binding motif enrichments in the consensus peaks were examined using Homer and i-cisTarget [127, 329]. For Homer, we extracted $\pm$150bp around the center of each consensus peak using BEDTools slop and the resulting FASTA files were used as input for HOMER findMotifs.pl (parameters -bits -mset vertebrates).
For i-cisTarget, genomic coordinates for the consensus peaks were mapped from the *mm10* to *mm9* reference genome using UCSC-liftOver. These were then uploaded to the i-cisTarget webserver to perform a full analysis with default settings (species = mm9) [329]. The resulting motif enrichments were downloaded, filtered for redundant motifs and visualised as heatmaps.

TF footprinting

TF binding motifs of interest were obtained from the JASPAR database [330]. We searched for motif matches using motifmatchr (*https://github.com/GreenleafLab/motifmatchr*) which implements the MOODS motif matching algorithm for R [331]. Searches were performed either genome-wide (for CTCF) or $\pm$5kbp around consensus peak summits (for CDX). The resulting matches were extended $\pm$150bp. An insertion count matrix at base pair resolution, centred on the motifs, was generated by counting adjusted Tn5 insertions. Only fragments smaller than 100bp were considered (nucleosome free fraction).
For the footprinting, we used PWM scores and corresponding insertion count matrix as the input for CENTIPEDE to compute posterior probabilities that a motif is bound [332].

A different threshold to the posterior probability was applied to classify bound/unbound depending on the motif matching strategy (genome-wide: $P \geq 0.99$); peak summit: $P \geq 0.9$).

Comparison of D5H and D5+

To compare the D5H and D5H+ conditions, we first identified robust peaks in D5H+ and added them to our consensus peak set. Differentially accessible peaks were identified using the same methods as above. Differential peaks were overlapped with the SOM clusters and used as input for i-cisTarget motif enrichment.

4.2.3 ChIP-seq analysis

Preprocessing

Sequencing adapters and low quality base calls were trimmed from reads using trim_galore with default settings. Trimmed reads were mapped to the *mm10* reference genome using bowtie2 with –sensitive as additional option. Alignments were filtered for unmapped, multi-mapping and duplicated reads. Signal tracks as log2 fold-change between ChIP and input samples were generated using deepTools bamCompare (parameters: –scaleFactorsMethod SES –ratio log2 -bs 25 –ignoreDuplicates). Peak calling was performed using MACS2 (parameters: -g mm -p 10^{-3}). Publicly available datasets were downloaded and analysed in the same manner (Supp. Table B.15.

Enrichment analysis

We checked if different groups of SOM classified consensus ATAC-seq peaks were enriched for TF binding using LOLA (default settings) [192] (Supp. Table B.16). We considered all TF ChIP-seq samples with an adjusted p-value$<$ 0.01 as enriched. To complement the LOLA *mm10* core database with TFs relevant for neural development, we added 4 newly generated samples and 39 publicly available TF ChIP-seq datasets (Supp. Table B.15. Peak calls from replicate ChIP-seq experiments were considered separately.

Gene ontology enrichment of CDX2 bound open chromatin sites

MNP CDX2 ChIP-seq peaks [333] were overlapped with NMP, NMP-SC, spinal cord, H/SC and hindbrain regions. The resulting peak sets were used as the input for gene ontology enrichment analysis using GREAT (default settings).

4.2.4 RNA-seq analysis

RNA-seq experiments in this study were quantified using Salmon (quasi-mapping mode) [154] with the GENCODE release M14. Both single-end and paired-end reads were processed using following options: -l A -seqBias -numBootstraps 50. The resulting counts and transcripts per million (TPM) were used for downstream analysis.
Previously published RNA-seq data from neural progenitor cells generated with the same differentiation protocol was downloaded from GEO or ArrayExpress and listed in Supp. Table B.13. Differential analysis of D5H versus D5SC samples was performed using DESeq2 with default settings.

4.2.5 Microarray analysis

Microarray data was downloaded from GEO and listed in Supp Table B.14. Preprocessing and normalisation was done with the robust multichip average algorithm (RMA) from the oligo package (default settings) [334]. Differential expression between two conditions was computed fitting a linear model with limma [281]. The resulting differentially expressed genes were annotated with gene symbols using the package mouse4302.db (*https://bioconductor.org/packages/release/data/annotation/html/mouse4302.db.html*).

4.2.6 Chromatin-Chromatin interaction analysis

A dataset of putative gene-region interactions was downloaded from the 4DGenome database [335] (Supp. Table B.16). The interactions were mapped to *mm9* and for further downstream analysis re-mapped to *mm10* using UCSC-liftOver with default settings. Interactions for which only one anchor could be mapped to *mm10* were removed. Putative chromatin-chromatin interactions were mapped by filtering for anchors which overlap open chromatin sites from this study.

4.3 Results

4.3.1 *In vitro* model of neural progenitor development generates cells with different AP-axis identities

Here we aim to improve our understanding of the spatiotemporal regulation of gene expression during the normal development of the nervous system along the anterior posterior axis (AP) [305]. To study this we used a previously described *in vitro* model system based on the differentiation of mouse embryonic stem cells (ESCs) to neural progenitor cells (NPCs) of different regional identities along the AP axis (Fig. 4.2A) [310, 311].

Briefly, ESCs were incubated for 2 days with FGF to generate epiblast-like cells (EpiLC). On day 3 either (i) sonic hedgehog (SHH) signalling was activated to generate a more anterior (A) forebrain like cell type or (ii) retinoic acid (RA) + SHH signalling were activated to promote differentiation to a hindbrain (H) like phenotype. Exposure to WNT signalling by treatment with a GSK3b inhibitor (CHIR99021) on day 2 initially differentiates EpiLC towards a bipotent progenitor cell fate called neuromesodermal progenitors (NMPs). Treatment of NMPs with RA + SHH induces differentiation towards a posterior spinal cord (SC) fate (Fig. 4.2A).

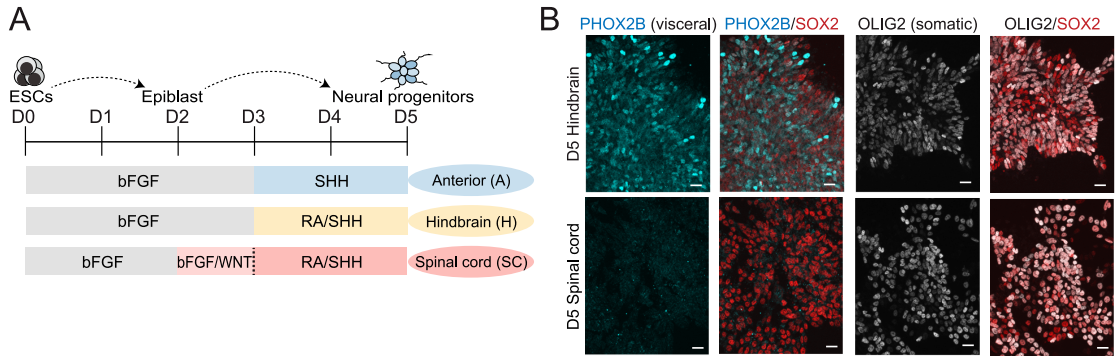


Fig. 4.2: *In vitro* differentiation of mESC to NPC of different regional identity along the AP axis.
(A) Schematic diagram illustrating the differentiation of mouse ESCs to NPCs with anterior (light blue), hindbrain (orange) and spinal cord (red) identities. Differentiation of spinal cord NPCs requires an intermediate NMP state (pink) which is generated by applying FGF and WNT signals. **(B)** Immunofluorescence staining of PHOX2B and OLIG2 in day 5 hindbrain and spinal cord NPCs. PHOX2B is a specific marker for visceral motor neurons and OLIG2 for somatic motor neurons (Scale bars = 20μm). Figure adapted from [305].

To confirm the cell identity of *in vitro* differentiated cells, we performed immunofluorescence imaging using the following marker genes: PHOX2B (visceral motor neuron progenitors (MNP)) and OLIG2 (somatic MNP; Fig. 4.2B) [336]. The choice of markers was based on the observation that the *in vivo* posterior spinal cord completely lacks any visceral MNPs. As expected, hindbrain NPCs were PHOX2B and OLIG2 positive. In contrast, spinal cord NPCs were OLIG2 positive but lacked PHOX2B expression.

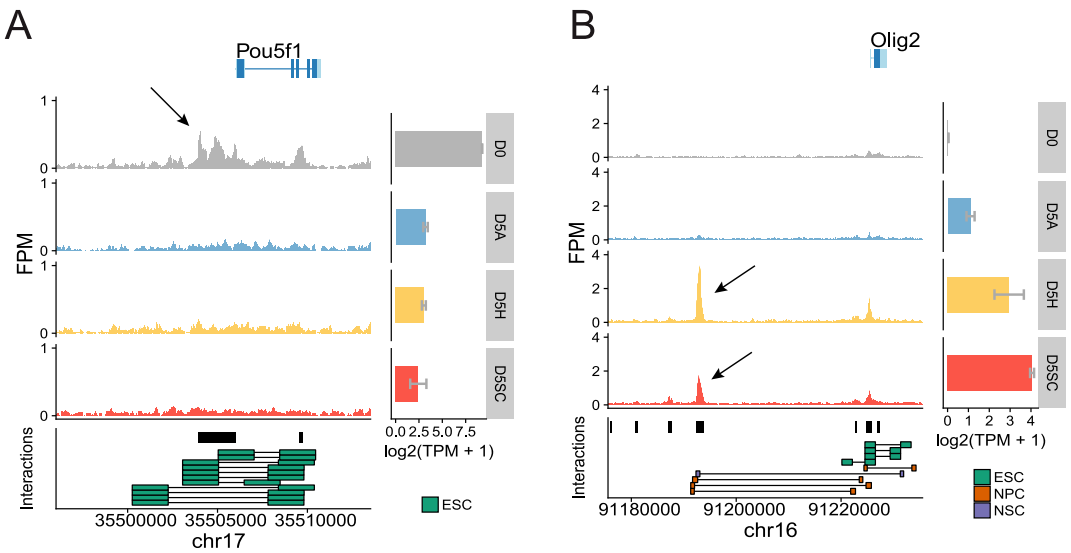


Fig. 4.3: Examples enhancers displaying changes in chromatin accessibilities during ESC to NPC differentiation.
(A) Genome browser views of ATAC-seq accessibility signal at the known ESC-specific *Pou5f1* enhancer [216]. ATAC-seq signal indicates an accessible enhancer at day 0 (arrow), but the signal is diminished on other days. Tracks include from top to bottom ESC (grey), day 5 anterior (blue), hindbrain (yellow) and spinal cord NPCs (red). Right panel shows barplots of *Pou5f1* or *Olig2* gene expression levels as TPM from mRNA-seq [310] (error bars = SEM). Lower panel shows chromatin-chromatin interactions from the 4DGenome database [335]. **(B)** The distal *Olig2* enhancer is accessible in day 5 hindbrain and spinal cord (arrows) [337].
Figure adapted from [305].

Next, we performed an initial assessment of chromatin accessibilities by ATAC-seq at each differentiation time point from day 0 ESC to day 5 anterior, hindbrain and spinal cord cells (Fig. 4.2A). Examining previously described enhancers confirmed the identity of generated cell types. For example, the proximal enhancer element of the *Pou5f1* gene is accessibile only on day 0, in line with the *Pou5f1* expression profile (Fig. 4.3A) [216]. In contrast, a known distal enhancer element of *Olig2* shows no accessibility on day 0 but specifically gains it on day 5 in hindbrain and spinal cord samples (Fig. 4.3B) [337]. *Olig2* is specifically

expressed in these day 5 NPCs. In both cases physical chromatin-chromatin interactions are found between the enhancer and gene (Fig. 4.3B) [335].
In summary, initial assessment of NPC identity via immunofluorescence of marker genes confirmed the successful differentiation of NPCs with different cell fates *in vitro*. ATAC-seq reveals cell-type specific changes in enhancer accessibilities during differentiation, suggesting that this assay will enable us to address our research question.

4.3.2 | ATAC-seq quality control analysis

Next, we performed quality control (QC) analysis of the ATAC-seq data to check that we obtained 2 sufficiently enriched replicates from each time point and perturbation condition (Fig. 4.4). ATAC-seq is based on the insertion of sequencing adaptors via a hyperactive Tn5 transposase into nucleosome free DNA [131]. A disadvantage of this is that mitochondrial DNA is completely nucleosome free and therefore usually enriched for Tn5 insertions. For this reason, we checked the proportion of fragments mapping to chrM which varied between 1.1% (D5H-plus rep2) and 34.2% (D3NMP-Cdx rep2; Fig. 4.4A). However, this is not necessarily a sign of bad quality as long as a there is sufficient fragment enrichment on the nuclear genome.
Another hallmark of ATAC-seq, if paired-end sequenced, is a periodic fragment length distribution (Fig 4.4B) [131]. As expected our samples showed enrichment of fragments smaller than 100bp originating from DNA binding proteins such as TFs. Further, fragments spanning mononucleosomes can be observed between 150bp and 247bp. This confirmed that our ATAC-seq experiments could recover expected fragment sizes.
It has previously been suggested to compute a transcriptional start site (TSS) enrichment score as a measure of Tn5 insertion enrichment quality [338]. The rationale behind this measure is that promoter regions offer a well-defined search space and clear nucleosome positioning compared with enhancers. Additionally, fragments in bulk ATAC-seq originate from heterogeneous cell populations where cells express genes from the same promoter but fine-tune expression using different long-range regulatory inputs. Initially for this analysis, the sum of insertions per base-pair around annotated TSS was computed, shown here as a metaprofile (Fig. 4.4D). Then the enrichment score was determined as the maximum at the TSS and divided by the background which was estimated via the minimum. We performed this analysis for all ATAC-seq samples leading to TSS scores between 3.3 (D2 rep2) and 12.3 (D5H rep2; Fig. 4.4C). Moreover, metaprofiles showed a second maximum at around 250bp downstream of the TSS corresponding to well positioned +1 nucleosomes (Fig. 4.4D) [338].
Next, we computed the fraction of fragments in consensus peaks (FRIP) to provide a

measurement using all detected open-chromatin regions (Fig. 4.4E). This was done since not all differentiation time points might have comparable number of accessible promoters. We observed fractions between 9.3% (D1 rep2) and 18.6% (D5H rep2) for our samples. The FRIP might be a more comparable measure than TSS enrichment between different conditions since it was computed for our consensus peak set. The TSS enrichment score might show higher variations due to potential differences in promoter/enhancer ratios between different conditions.

Last, we assessed the possibility of TF footprinting with our ATAC-seq data. For this we first matched CTCF motifs to the mouse genome and computed an insertion per bp matrix around motif centres. This was used as input for CENTIPEDE, a framework based on an unsupervised Bayesian mixture model to distinguish bound from unbound motifs based on insertion patterns [332]. Classified motifs were used to compute the enrichment of insertions around bound sites over background (unbound CTCF motifs; Fig. 4.4F). Using this approach, we could recover the expected insertion profiles as previously described by Buenrostro *et al.* Thus, confirming the sufficient resolution of our data to look at other TF footprints [131] .

Taken together, our QC analysis confirmed the sufficient insertion enrichment and resolution of our ATAC-seq samples for the following downstream analysis.

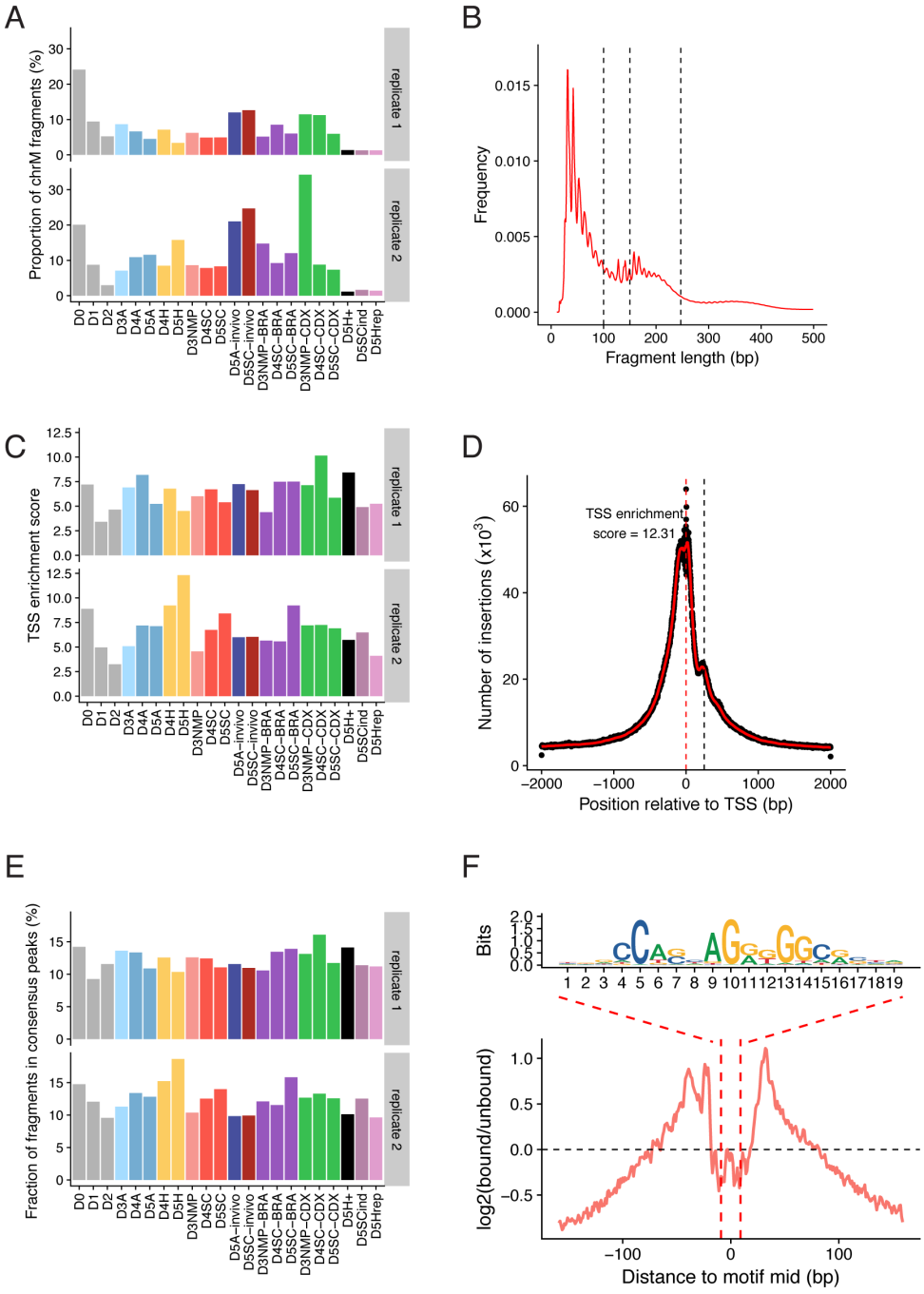


Fig. 4.4: Quality control analysis for all in-house generated ATAC-seq samples.
(A) Percentage of fragments mapping to chrM for each condition and both replicates. Colour denotes experimental condition. **(B)** Representative fragment length distribution for day 0 replicate 1. Dashed lines represent nucleosome free fragments smaller than 100bp and mononucleosome fragments between 150bp and 247bp. **(C)** Barplots summarising TSS enrichment scores for each condition and both replicates. Colour as in (A). **(D)** Tn5 insertion profile ±2kb around annotated TSS sites. The sample with the highest TSS enrichment score was chosen (D5H replicate 2). The red dashed line marks the TSS and the black line the position of the +1 nucleosome. Smoothed metaprofile is shown as a red line and was computed as a 50bp running average. **(E)** Percentage of fragments in consensus open-chromatin peaks for each condition and both replicates. Colour as in (A). **(F)** Tn5 insertion profile ±150bp around the CTCF consensus motif. Shown sample corresponds to day 0 replicate 1. CTCF consensus motif plotted on top was taken from JASPAR 2018 [330]. Figure adapted from [305].

4.3.3 Differential chromatin accessibility defines neural cell fates

First, we created a consensus peak set by peak-calling on each biological replicate independently, then computing a reproducible consensus set between replicates via IDR thresholding and finally merging the reproducible peaks per condition to a final set resulting in 78,223 consensus peaks (see Methods).

We used this consensus peak set to perform a more systematic analysis of differential accessibility by comparing every time point against day 0 to generate a set of regions with dynamic accessibility (Fig. 4.5A). For example, comparing day 5 SC with day 0 revealed that 13.9% (10,880/78,223) of consensus peaks have statistically significantly higher accessibility and 17.6% (13,804/78,223) have decreased accessibility ($FDR < 0.01$ and $|log2(FC)| > 1$). Moreover, this showed an increasing number of newly accessible regions with time during differentiation (Fig. 4.5B). However, differences in the total numbers of gained peaks were observed, with anterior NPCs gaining the fewest and hindbrain NPCs the highest numbers of accessible sites.

To confirm that the observed open-chromatin regions correspond to neural specific regulatory elements, we used a previously published H3K27ac ChIP-seq dataset for NPCs differentiated from ESCs, using a similar differentiation protocol [339]. H3K27ac is a histone modification enriched at active regulatory regions including promoters as well as enhancers [83]. We observed a depletion on ESC specific sites and an enrichment for neural specific regions (Fig. 4.5C).

This illustrates the chromatin dynamics during the differentiation from ESCs to NPCs and confirms that cells acquire a neural identity that is distinct to the pluripotent state [340, 310].

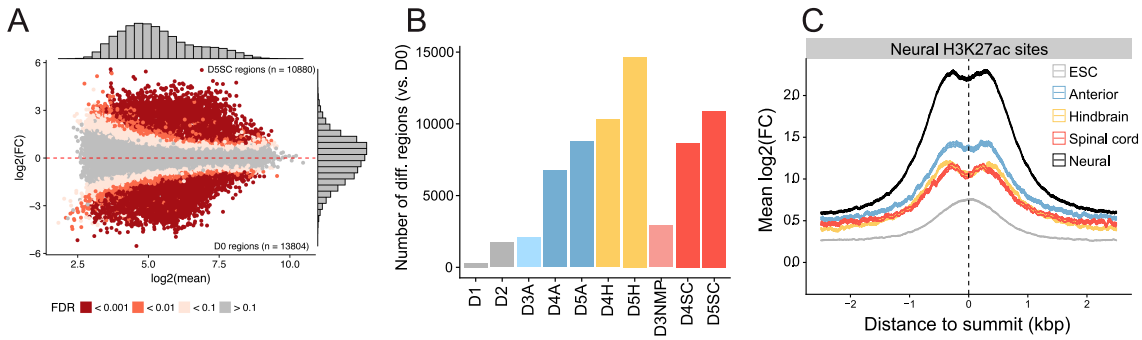


Fig. 4.5: Genome-wide differential analysis of ATAC-seq accessibility uncovers dynamic sites. **(A)** MA-plot showing differential accessibility analysis of open-chromatin peaks between day 5 SC and day 0. Colour denotes FDR, x and y-axis histograms summarise the distribution of log2 mean values and log2 fold-changes values respectively. **(B)** Number of differential accessible sites from comparison between each condition and day 0. Number of regions after thresholding for $FDR < 0.01$ and $log2(FC) > 1$. Colour marks experimental condition. **(C)** Metaprofile of H3K27ac ChIP-seq signal enrichment ± 3 kb around peak summits of ESC, anterior, hindbrain, spinal cord and neural regulatory regions. ChIP-seq is normalised as log2 fold-change over input. Figure adapted from [305].

To resolve the temporal dynamics of chromatin accessibility during NPC differentiation, we clustered all peaks with significantly increased accessibility using a 5x5 self-organising map (SOM; Fig. 4.6A) [29]. Resulting SOM clusters were merged based on similar accessibility trends and annotated according to the condition displaying the highest accessibility. Assigned cluster annotations were confirmed by manual examination to check overlaps with previously described regulatory regions (e.g. Fig. 4.3). This revealed nine biological meaningful clusters: early differentiation (Epi, NMP, NMP/SC), a larger cluster specific for neural cell fate (Neural), 3 clusters specific for NPC regional identity (A/H/SC), one shared cluster between hindbrain and spinal cord and one shared cluster between hindbrain and anterior (Fig. 4.6A, A").

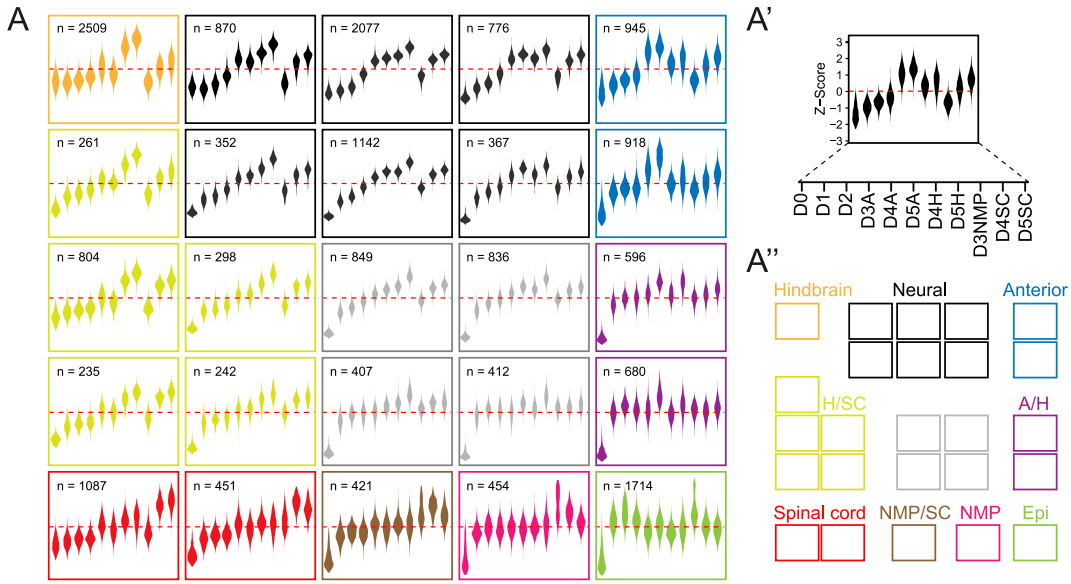


Fig. 4.6: Clustering of differential accessible sites using a self-organising map uncovered chromatin dynamics.
(A) Clustering of differential accessible regions using a 5x5 SOM. Each cell illustrates underlying dynamics as a violin plot of Z-scores derived from chromatin accessibility in different conditions. **(A')** Zoom in on an example cell from showing chromatin-accessibility Z-scores from 11 ATAC-seq conditions representing different time points in differentiation. **(A'')** Map illustrating the manual clustering of SOM cells into biological clusters. These include clusters representing early differentiation stages (prior day 4) Epi ($n = 1,714$; green), NMP ($n = 454$; pink) and NMP/SC ($n = 421$; brown). Three clusters corresponding to AP-axis identity; anterior ($n = 1,863$; blue), hindbrain ($n = 2,509$; orange) and spinal cord ($n = 1538$; red). Two clusters showing shared regions between anterior/hindbrain ($n = 1,276$; purple) and hindbrain/spinal cord ($n = 1,840$; lime). A large common cluster for neural specific sites ($n = 5,584$; black) and another unassigned cluster ($n = 2,504$; grey). Figure adapted from [305].

Clusters corresponding to regions specific for AP-axis regional identity were confirmed by the recovery of previously described enhancers. For example, the anterior cluster included an enhancer located within the second intron of *Shh* (Fig. 4.7A). The same regulatory region was previously described to drive *Shh* expression in the developing midbrain [341]. In contrast, we observed an enhancer distal to *Phox2b* showing accessibility only in hindbrain (Fig. 4.7B). This is in line with *in vivo* expression of *Phox2b* in the hindbrain and its importance for normal hindbrain development [336, 342]. Spinal cord specific regulatory regions overlapped with a known enhancer flanking the posterior expressed *Hoxc8* gene (Fig. 4.7C) [343]. Together, these analyses identified biologically relevant chromatin dynamics that captures ESC differentiation into NPC with different regional identities.

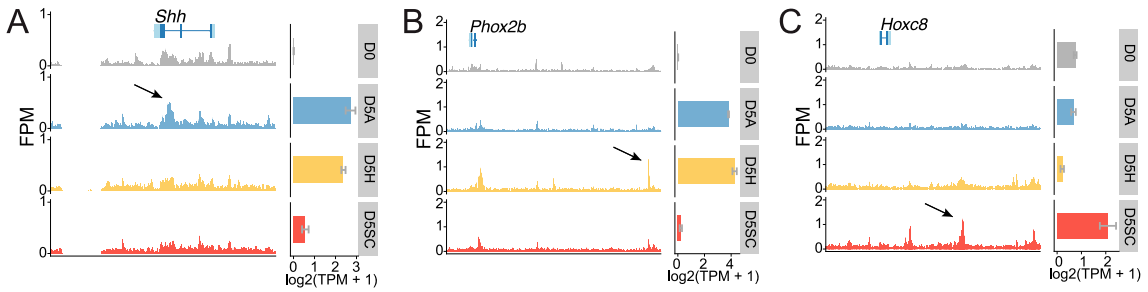


Fig. 4.7: Examples for enhancer elements specific for NPC AP-axis regional identity.
(A) Example of an anterior NPC specific enhancer region located within the second intron of *Shh* [341]. ATAC-seq is shown as FPM for day 0 (grey), day 5 A (blue), day 5 H (yellow) and d 5 SC (red) conditions. Right panel shows barplots summarising average gene expression as log2(TPM+1) from mRNA-seq (error bars = SEM from 3 replicates) [310]. **(B)** As in (A) but for an enhancer distal to *Phox2b* which is specific to hindbrain NPCs. **(C)** As in (A) but for a *Hoxc8* enhancer which is specifically accessible in spinal cord. Figure adapted from [305].

4.3.4 Regions of dynamic accessibility correspond to enhancer elements

Annotation of regions relative to gene features was performed to gain a better understanding which genomic elements are involved in the differentiation process. AP-axis identity clusters (A/H/SC) showed that a majority of regions were located either in distal intergenic (2,474/5,910, 41.9%) or intronic regions (2,421/5,910, 41%; Fig. 4.8A). In contrast, shared neural sites are mostly found at gene promoters (3,803/5,584, 61.8%; Fig. 4.8A). To check if these regions are active in NPCs, we again checked for enrichment of H3K27ac ChIP-seq in the different clusters [339]. Strong enrichment was observed for neural sites corresponding to actively transcribed promoter regions in NPCs (Fig. 4.5C). AP-axis regional identity enhancers (A/H/SC) showed enrichment to a lesser extent than neural sites but still pointing towards their activity in neural cells. Further evidence for their neural enhancer activity was found by overlaying them with the VISTA database containing *in vivo* validated developmental enhancers [344]. A significant enrichment of fore-/mid- and hindbrain active enhancers were observed in anterior, hindbrain and neural sites ($p < 10^{-16}$, one-sided Binomial test), but not for epiblast nor spinal cord sites (Fig. 4.8B). However, spinal cord sites were significantly enriched for enhancers showing activity in the neural tube as well as limbs ($p < 10^{-16}$, one-sided Binomial test). In addition, we checked the overlap of our enhancers with DHS sites derived from various tissues as well as developmental time points [345, 327]. This analysis revealed that more than 80% of sites in hindbrain/spinal cord cluster could be found in open-chromatin regions from whole embryo samples (Fig. 4.8C, D). Additionally, for both examined clusters we observed a high concordance with DHS from developing neural

tissues (∼60%-80% peak overlap). Taken together, these analyses revealed that regions in the AP-axis regional identity clusters correspond to active developmental enhancer elements and regions shared between all NPCs (neural sites) to promoter regions. This is in line with the view that, during development, cell-type specific enhancer elements regulate expression levels of common sets of genes.

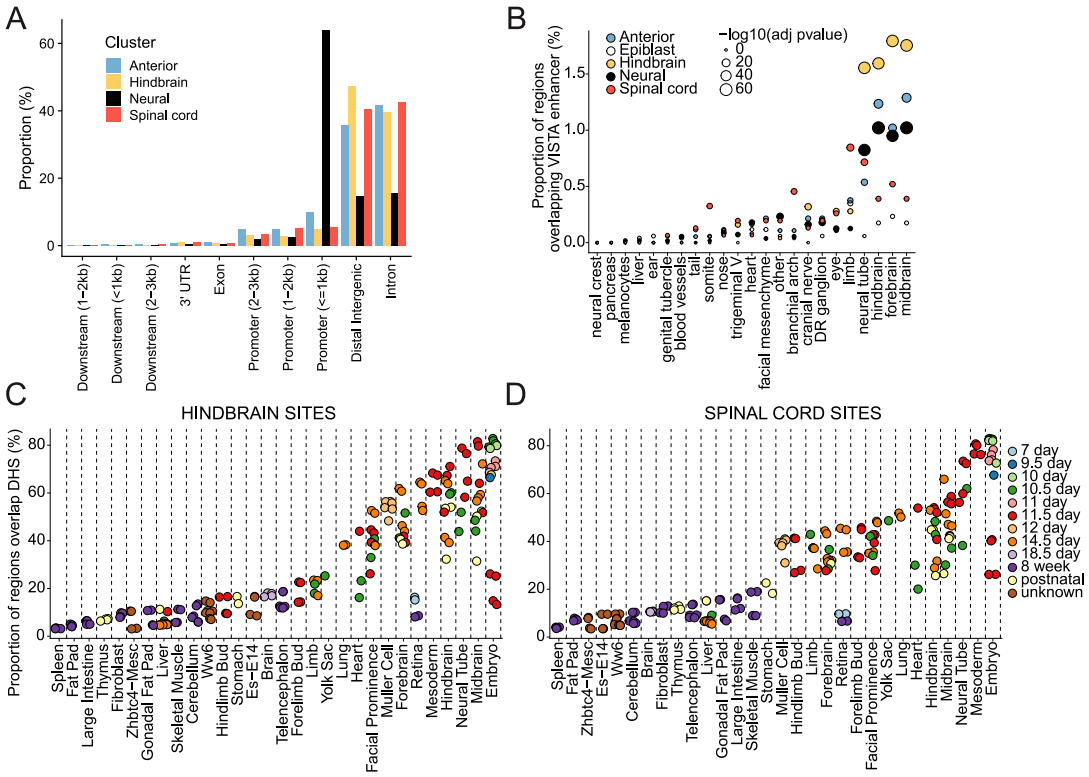


Fig. 4.8: Overlap of neural and AP regional identity clusters with genome annotation.
(A) Proportion of regions belonging to neural and AP-identity (A/H/SC) clusters overlapping with different genomic features. **(B)** Enrichment and overlap of VISTA enhancers for epiblast, neural and AP regional identity (A/H/SC) region sets (p-value: one-sided Binomial test; encoded as circle size). **(C)** Overlap of hindbrain-specific regulatory elements with published DNase hypersensitive sites (DHS) from various different tissues and time points during development. **(D)** Overlap of spinal cord specific sites with published DHS from various different tissues and time points during development. Figure adapted from [305].

4.3.5 *In vitro* enhancer signatures exhibit similar accessibility dynamics *in vivo*

Next, we examined whether our *in vitro* enhancer signatures show similar chromatin accessibility during *in vivo* NPCs development. Importantly, we wished to validate that our

NPC specific enhancer signatures (A/H/SC) reflect *in vivo* accessibility differences along the AP-axis of the neural tube. E9.5 mouse embryos were dissected into the two ends of the anterior neural tube (up to the hindbrain) and spinal cord (mostly thoracic) tissues. To isolate NPCs, subdissected tissues from *Sox2eGFP* transgenic embryos, that express GFP throughout the nervous system, were used together with FACS [311]. We performed ATAC-seq on the resulting *in vivo* anterior and spinal cord NPCs that were isolated in the GFP positive fraction, which are comparable to *in vitro* day 5 cells (Fig. 4.9A) [311]. Initially assessment of the *in vivo* ATAC-seq samples revealed chromatin accessibility at well-known enhancers such as the anterior specific intronic *Shh* enhancer (Fig. 4.9B) and a distal enhancer of *Olig2* involved in spinal cord development (Fig. 4.9C) [341, 337].

Next, we overlapped peaks called on the *in vivo* ATAC-seq samples with the SOM clustered *in vitro* peaks. This showed significant enrichment of spinal cord enhancers in the *in vivo* spinal cord peaks compared with the peak set from anterior neuroectoderm (anterior: 73.1% (1,362/1,863); spinal cord: 59.7% (1113/1863); $p = 3.73 \cdot 10^{-21}$, one-sided Fisher's exact test). In contrast, peaks from *in vivo* anterior samples displayed higher enrichment of anterior enhancers compared with the spinal cord sample (anterior: 44.1% (679/1,538); spinal cord: 63.8% (981/1,538);$p = 4.28 \cdot 10^{-28}$, one-sided Fisher's exact test).

Genome-wide comparison of accessibility changes between spinal cord/anterior *in vivo* and *in vitro* datasets confirmed the similarity of chromatin changes (Fig. 4.9D). Additionally, anterior and spinal cord specific enhancer signatures showed statistically significantly accessibility differences for both *in vitro* as well as *in vivo* samples ($p < 10^{-16}$, two-sided Wilcoxon test; Fig. 4.9D, E). However, our *in vivo* ATAC-seq showed on average poorer Tn5 insertion enrichment and higher background noise (Fig. 4.9B, C, E). This could be explained by the lower cell numbers and higher degree of cellular heterogeneity in the *in vivo* samples [346]. In summary, the comparison of *in vitro* and *in vivo* NPCs validated the concordance of observed accessibility changes along the AP-axis and supports the specificity of extracted enhancer signatures.

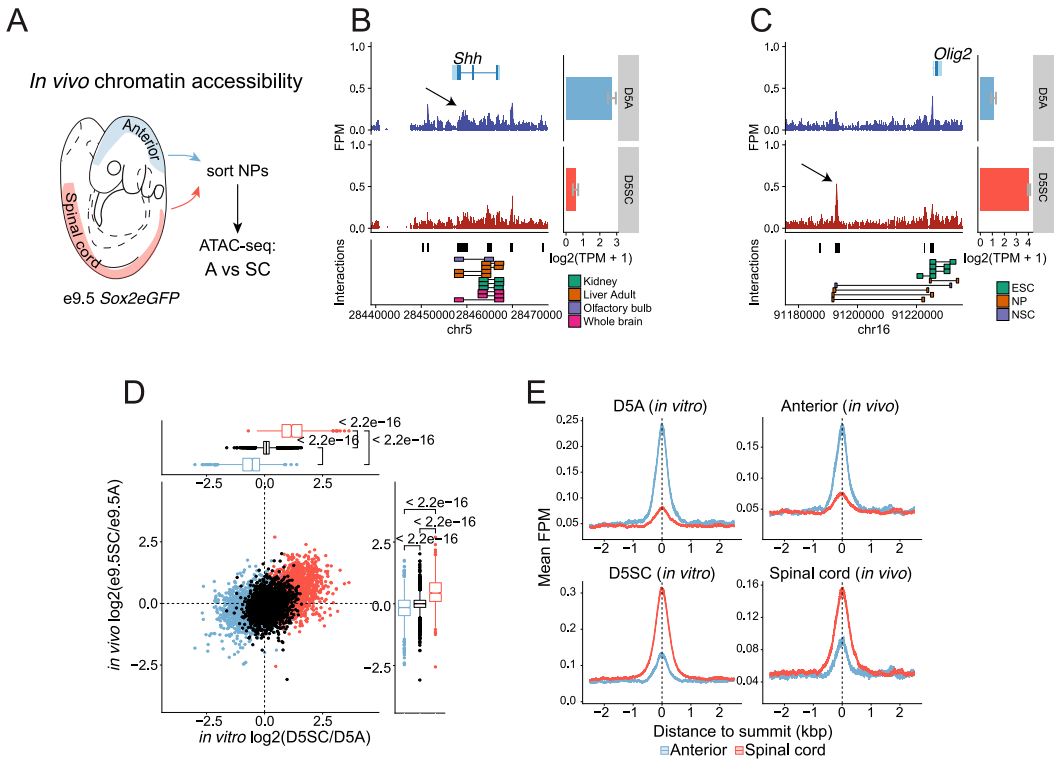


Fig. 4.9: Comparison of ATAC-seq derived from *in vitro* and *in vivo* NPCs.
(A) Schematic illustrating workflow to determine *in vivo* chromatin accessibility. E9.5 mouse embryos were dissected and *Sox2eGFP* expressing cells FACS sorted. Obtained cell populations were used for ATAC-seq. **(B)** Anterior specific *Shh* enhancer shows accessibility *in vitro* [341]. *Right:* Barplot shows expression from *in vitro* RNA-seq (Error bar = SEM). *Bottom:* Plot shows chromatin-chromatin interactions from 4DGenome database. Colour shows experimental condition. **(C)**. Also, in (B) but for spinal cord specific enhancer located upstream of *Olig2*, which shows accessibility *in vitro*. **(D)** Scatterplot comparison of accessibility in *in vitro* day 5 anterior vs. spinal cord samples against *in vivo* E9.5 anterior vs. spinal cord samples. Accessibility changes were computed for following regulatory region signatures: anterior (blue), neural (black) and spinal cord (red) regulatory regions (p-values from two-sided Wilcoxon test). **(E)** Metaprofiles showing accessibility of *in vivo* and *in vitro* enriched anterior (blue) and spinal cord (red) enhancers. Figure adapted from [305].

4.3.6 Motif and ChIP-seq enrichment analysis uncovered major regulators of observable chromatin dynamics

Next, we asked which TFs might be responsible for establishing the observed chromatin accessibilities in the different NPCs. To answer this, we performed TF motif as well as ChIP-seq peak enrichment analyses. We performed SOX2 ChIP-seq in day 5 hindbrain and spinal cord NPCs. In addition, 30 publicly available TF ChIP-seq datasets for relevant neural cell types were downloaded and processed (Supp. Table B.15). Resulting TF peak sets were

used for enrichment analysis with LOLA [192].

Enhancers specific for an anterior identity showed significant enrichment for FOXA2, NKX2.2, SOX2 and PTF1A binding sites ($FDR < 0.01$, one-sided Fisher's exact test; Fig. 4.10A). These observations were further supported by enrichments of Forkhead (e.g. FOXA1) and Sox family motifs in the same enhancer set (Fig. 4.10D, G). Moreover, *Foxa1* and *Foxa2* showed also more specific expression in day 5 anterior NPC (Fig. 4.10D). These enriched TFs were in line with their previously published function during anterior nervous-system development for example FOXA2 was described to be involved in midbrain development [347]. Furthermore, NKX2.2 was associated with forebrain organisation in *X. laevis* and PTF1A with induction of neural stem cells from fibroblasts [348, 349].
In contrast, hindbrain enhancers were enriched for factors like OLIG2, NKX6.1, SOX2, NKX2.2 and SOX3 ($FDR < 0.01$, one-sided Fisher's exact test; Fig. (4.10B). Motifs of Sox family members (e.g. SOX3, SOX10, SOX2) were most highly enriched and *Sox3* is showing the most specific gene expression in hindbrain NPCs (Fig. (4.10E, H). Observed regulators were additionally supported by published literature for example OLIG2, NKX6.1 and NKX2.2 are involved in the establishment of motor neuron identity in the hindbrain [350, 351]. SOX3 was previously suggested as an early regulator of neuroectoderm development in zebrafish [352].
Finally, spinal cord elements showed statistically significant enrichment for CDX2, HOXC9, OLIG2 and SOX2 ($FDR < 0.01$, one-sided Fisher's exact test; Fig. 4.10C). We found further evidence for the importance of these factors by their high motif abundance in combination with highly specific expression of *Cdx* and *Hox9* genes in day 5 spinal cord NPCs (Fig. 4.10F, I). These observations are supported by the published findings; for example CDX2 is involved in spinal cord NPC differentiation and CDX null mice fail to develop a posterior axis [315, 353, 311]. OLIG2 and HOXC9 were both shown to be associated with motor neuron development in the spinal cord [354, 355, 356].
Further, SOX2 binding sites were enriched in hindbrain or spinal cord according to their cell type of origin, emphasising differential binding of factors expressed in both conditions (Fig. 4.10B, C). This again is in-line with previously publications showing wide-spread *Sox* expression during nervous system development along the AP-axis and involvement in the development of various neural cell types [357, 322].
Taken together, we were able to recover TFs known for their roles in neural development and we uncovered their putative roles as regulators involved in establishing the chromatin landscape in NPC along the AP-axis. Furthermore, the enrichment of tissue specific SOX2

ChIP-seq peaks, which is broadly expressed in the neural tube, confirmed again the specificity of the regional identity signatures from the SOM.

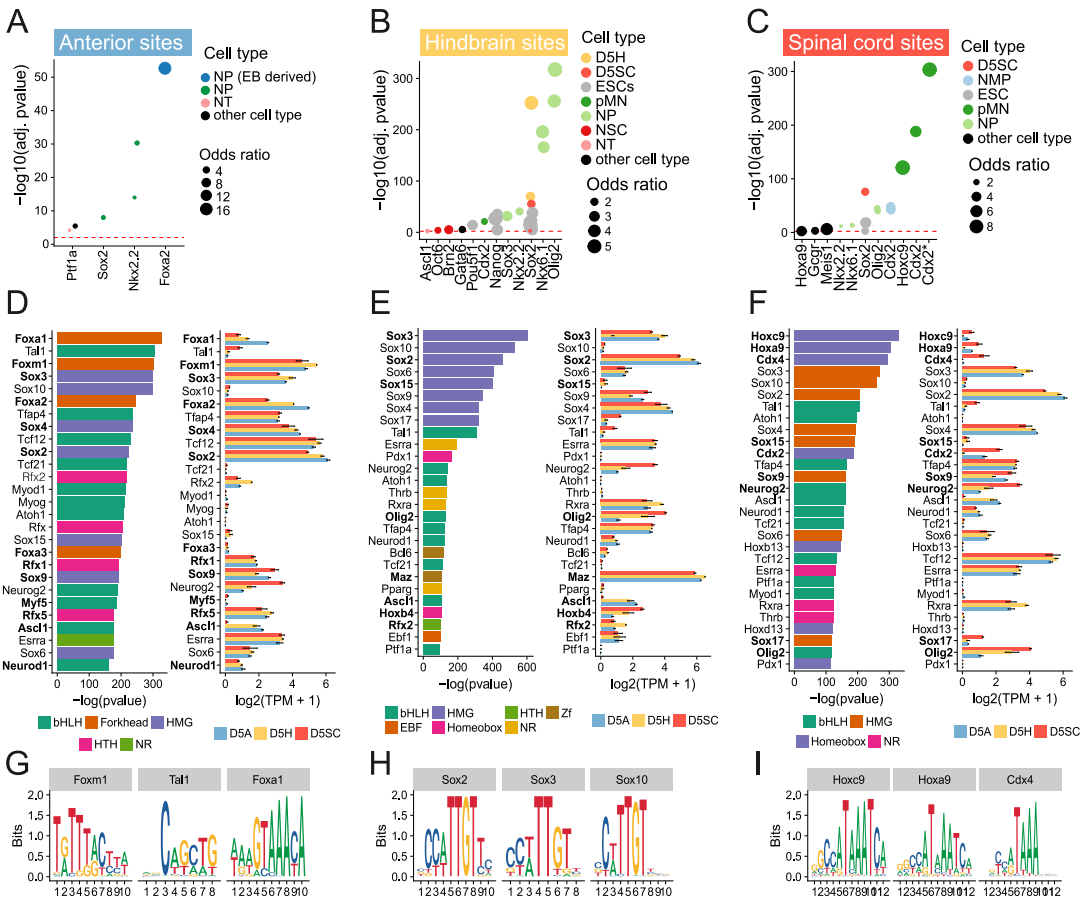


Fig. 4.10: Enrichment of TF motifs and ChIP-seq peaks in regulatory elements defining AP-axis regional identity.

(A-C) ChIP-seq peaks enrichment analysis using LOLA (p-values: Fisher's exact test). TF enrichment in sets of regulatory regions defining (A) anterior, (B) hindbrain and (C) spinal cord identity. Cdx2* refers to FGF treated somatic motor neuron progenitors (pMN). **(D-F)** Motif enrichment in sets of regulatory regions defining (D) anterior, (E) hindbrain and (F) spinal cord. *Left:* Plot showing unadjusted p-values for top enriched motifs in each condition (p-values: binomial test). *Right:* Barplot summarising expression of the TFs associated with the motif in each condition at day 5 A/H/SC (Error bar: SEM). Differentially expressed TFs between Anterior/Hindbrain/Spinal cord are highlighted by bold gene symbols ($FDR < 0.05$, DESeq2). **(G-I)** Sequence logos showing motifs associated with the top 3 TFs enriched in each region set. Bits refers to the information content at each position. Figure adapted from [305].

4.3.7 Chromatin dynamics induced by WNT depends on developmental time

Interestingly we could identify two clusters of transient regulatory regions, which gain accessibility upon exposure to WNT signalling on day 2 (Fig. 4.6A, NMP and NMP/SC). Previous studies have emphasised the importance of a related cell population, called neuromesodermal progenitors (NMPs), in mesoderm as well as spinal cord development [310, 311, 318]. We hypothesised that NMPs should initially commit towards a posterior identity and then acquire a neural cell fate. To test this, we took advantage of our *in vitro* model system and delayed induction of WNT signalling from days 2 to day 4 following RA treatment (Fig. 4.11A). The new condition was called hindbrain-plus since RA treatment normally induces NPC with hindbrain identity. This raised the question of whether hindbrain-plus NPC acquire spinal cord or hindbrain identity; would it be possible to delay WNT signalling and still produce spinal cord NPC?

Profiling of different marker genes via qRT-PCR revealed Notum expression in day 5 hindbrain-plus which confirmed successful induction of WNT signalling (Fig. 4.11B). However, hindbrain-plus cells failed to express posterior *Hox* genes such as *Hoxb9* and *Hoxc8* suggesting that these cells fail to induce spinal cord identity (Fig. 4.11C, D). To check if these cells would rather differentiated to a NMP like state first, we profiled the NMP markers *T/Bra* and *Cdx2* which were not expressed either (Fig. 4.11E, F). In contrast, hindbrain-plus cells failed to express hindbrain/motor neuron progenitor markers like *Phox2b* and *Olig2* (Fig. 4.11G, H). However, these NPCs started to express dorsal and intermediate neural tube markers such as *Pax7* and *Dbx1* (Fig. 4.11I, J) [358]. Thus, hindbrain-plus NPCs failed to acquire a posterior identity indicating that the timing of WNT induction on day 3 appears to be crucial.

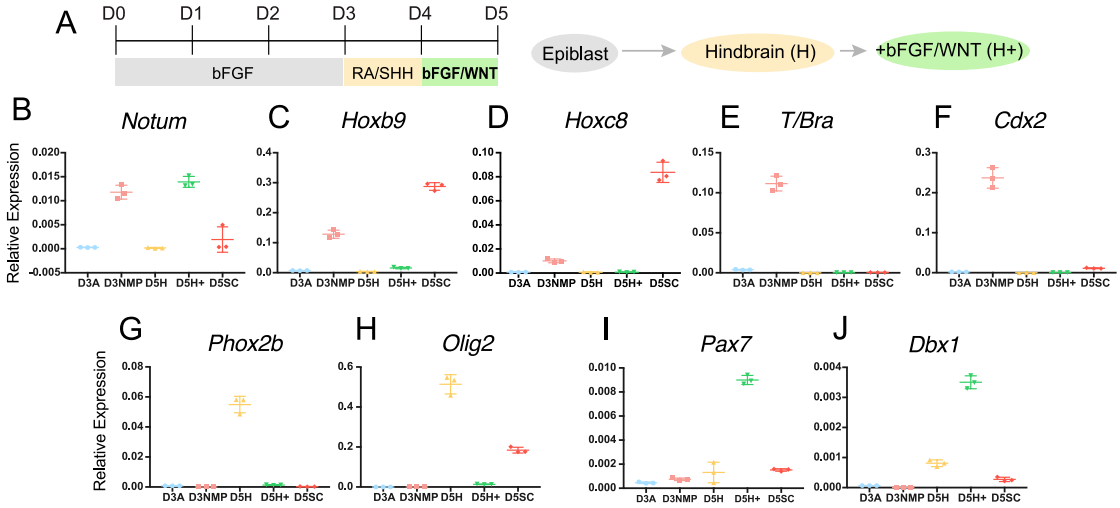


Fig. 4.11: Gene expression profiling on NPC with delayed WNT induction.
(A) Schematic illustrating the differentiation to day 5 hindbrain followed by activation of bFGF and WNT signalling resulting in hindbrain-plus NPCs. **B-J** Expression of marker genes measured by qRT-PCR in 5 different conditions including day 3 anterior/NMP and day 5 hindbrain/hindbrain-plus/spin cord. Expression is quantified as relative to β-actin. Figure adapted from [305].

To confirm these observations on a genome-wide level, we performed ATAC-seq on day 5 hindbrain-plus NPCs. Resulting ATAC-seq samples passed our quality control and peak calling yielded 23,156 consensus peaks between replicates (Fig. 4.4). Initial comparison of hindbrain/spinal cord enhancer accessibility in day 5 NPCs revealed that day 5 hindbrain-plus retained a more anterior cell fate that is more comparable to hindbrain than the spinal cord (Fig. 4.12A). Further, differential accessibility analysis between day 5 hindbrain and hindbrain plus showed that most regions increasing accessibility upon WNT induction are specific to hindbrain-plus identity (Fig. 4.12B). In contrast, the regions losing accessibility overlapped most frequently with hindbrain, H-SC and neural sites. Thus, indicating that the hindbrain-plus cells were not comparable with posterior spinal cord or ventral hindbrain NPCs. These observations, taken together with *Pax7/Dbx1* marker expression, pointed more towards hindbrain NPC with a more dorsal identity [359, 358]. Moreover, confirming that open-chromatin profiling by ATAC-seq was able to distinguish cells with different regional identities along the anterior-posterior, as well as dorsal-ventral, axis. Next, to assess which factors orchestrate the chromatin dynamics on day 5 upon activation of WNT signalling, we performed motif enrichment analysis. Regulatory regions opening up in day 5 hindbrain-plus showed enrichment of TCF/LEF TF motifs, indicative of canonical WNT signalling (Fig. 4.12 C). Bye contrast to day 5 hindbrain accessible regions which were enriched for

motifs associated with the SOX family.

Taken together, we observed that activation of WNT signalling on day 5 failed to induce a posterior spinal cord identity in NPCs. Therefore, the timing of WNT signalling is crucial for establishing the initial posterior identity which can then give rise to a posterior spinal cord neural cell fate.

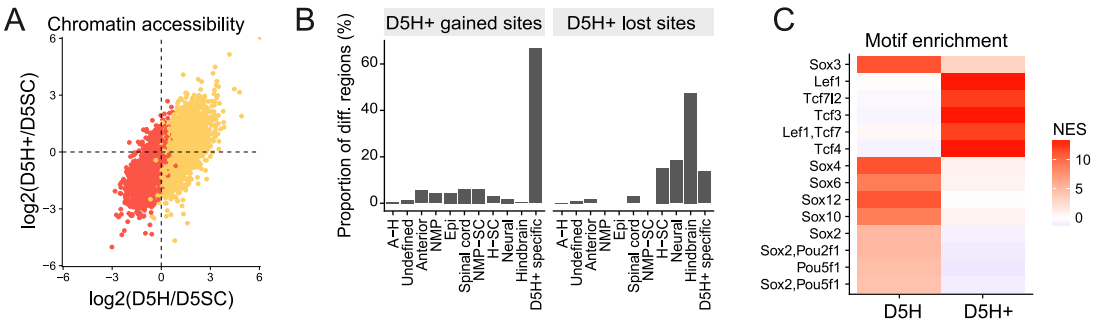


Fig. 4.12: ATAC-seq analysis of day 5 hindbrain-plus NPCs.
(A) Comparison between log2 accessibility ratios of day 5 hindbrain/spinal cord against hindbrain-plus/spinal cord within enhancer elements falling either into hindbrain (orange) or spinal cord (spinal cord) SOM cluster. **(B)** Proportion of differential accessible sites between day 5 hindbrain and hindbrain-plus overlapping SOM clusters. **(C)** iCisTarget motif enrichment for differential accessible sites between day 5 hindbrain and hindbrain-plus. Colour denotes Normalised enrichment scores (NES) computed with iCisTarget. Figure adapted from [305].

4.3.8 NMPs show transient chromatin accessibility induced by WNT signalling

Since hindbrain-plus cells were neither comparable with NMP nor spinal cord identities, we examined if WNT signalling on day 3 induced the NMP state already primed towards a posterior identity. Initially, we compared our identified *in vitro* NMP enhancer signature with accessibility peaks from in-house as well as publicly available *in vivo* ATAC-seq to ensure their relevance in an *in vivo* context [315, 325]. This confirmed good agreement of our NMP enhancers with open-chromatin regions from E7.5 posterior tissue, which normally harbours the NMP population (Fig. 4.14A). In contrast, there is very little overlap with regions from E9.5 samples ($< 10\%$), again in agreement with the concept of NMPs harbouring a genomic identity that is distinct to neural cells.

Having confirmed the existence of our NMP signature *in vivo*, we performed ChIP-seq enrichment analysis to identify TFs involved in WNT signalling integration and enabling the observed transient accessibility. Our analysis revealed enrichment of CDX2 ChIP-seq peaks, from epiblast-derived stem cells (EpiSC) as well as induction in visceral MNP, in our NMP enhancer signature (Fig. 4.13A) [333, 315]. This observation was in line with previous gene expression studies of *in vitro* differentiated and *in vivo* isolated NMPs highlighting the importance of *T/Bra*, *Sox2* and *Cdx* for this cell state [310, 311, 318]. Additionally, our previous analysis showed that spinal cord sites are highly enriched for CDX2 motif and ChIP-seq peaks (Fig. 4.10B, F). To confirm CDX2 binding to our NMP enhancer set under *in vitro* conditions, we performed TF footprinting around CDX2 binding motifs using again CENTIPEDE [332]. A clear enrichment of Tn5 insertions around CDX2 motifs classified as bound was observable in day 3 NMP ATAC-seq (Fig. 4.13B *left panel*). In contrast, CDX2 motifs classified as unbound showed clear depletion of Tn5 insertions. To validate the observed CDX2 footprint, we made use of CDX KO ATCA-seq under similar conditions. Visualising the insertion frequency over previously classified bound/unbound CDX2 motifs showed depletion of insertions around both motif classes and therefore confirms the causality between CDX binding and observed insertion pattern (Fig. 4.13B *right panel*). Altogether this data as well as previous published literature emphasises the importance of CDX factors in establishing chromatin accessibility upon WNT signalling for the intermediate NMP state [315, 325].

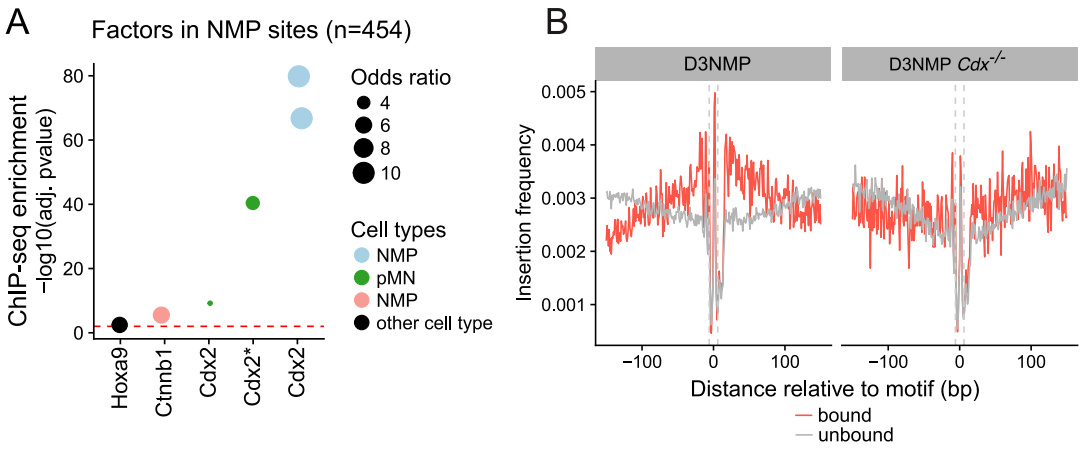


Fig. 4.13: NMP enhancers are enriched for CDX2 binding sites.
(A) Enrichment analysis of TF ChIP-seq peaks in NMP enhancers (p-value: one-sided Fisher's exact test). **(B)** Tn5 insertion profiles relative to CDX2 consensus motifs associated with NMP sites for day 3 NMP wild-type and *Cdx* KO cells. Motifs were discriminated into bound (red) and unbound (grey) sites using CENTIPEDE [332]. Figure adapted from [305].

4.3.9 CDX2 is the master regulator of NMP specific regulatory elements

Taking together all previously published evidence and our observations of CDX2 enrichment as well as binding to NMP enhancers, we hypothesised that CDX factors play a major role in establishing the transient open chromatin landscape necessary for NMP identity [315, 325]. To test this assumption, we used an ESC cell line containing either *Cdx1/2/4* triple knockout (KO) or *T/Bra* KO. T/BRA was highlighted by previous studies for its importance in paraxial mesoderm development from NMPs and therefore serves as negative control [311, 318]. The KO ESCs were then exposed to the same signals, which normally are sufficient to differentiate wild type (WT) ESCs to spinal cord NPCs. Again, we measured chromatin accessibility under *Cdx* and *T/Bra* KO conditions on days 3, 4 and 5 using ATAC-seq. All replicates of the KO ATAC-seq data sets passed our quality control (Fig. 4.4). Peak calling on these samples resulted in consensus peaks sets of following sizes: *T/Bra* KO D3NMP 30,911; D4SC 29,220; D5SC 36,111 and *Cdx* KO D3NMP 28,204; D4SC 27,934; D5SC 24,936.

Initial analysis of *T/Bra* KO mutants revealed that these cells have comparable accessibilities during the entire differentiation process and successfully generate spinal cord NPCs (Fig. 4.14B, C). Further quantification of enhancer accessibility showed no significant difference between NPCs differentiated from WT or KO ESCs (NMP: $p = 0.5$, NMP-SC: $p = 0.68$,

two-sided Wilcoxon test; Fig. 4.15A, B, C). These observations were in line with previous observations that T/BRA is essential for paraxial mesoderm development [311, 318].

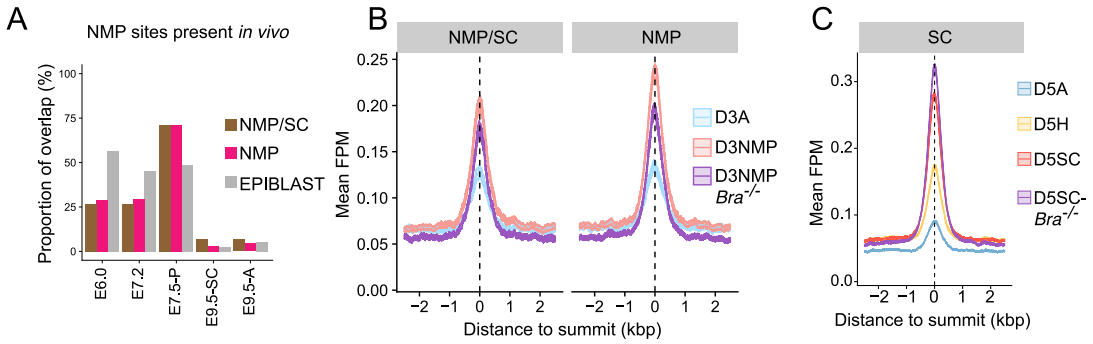


Fig. 4.14: NMP enhancer accessibility is not regulated by T/BRA.
(A) Overlap of epiblast, NMP and NMP-SC enhancers with open chromatin sites from published *in vivo* ATAC-seq. **(B)** Metaprofile illustrating mean accessibility, measured by ATAC-seq, within $\pm 2.5kb$ around NMP and NMP-SC enhancer summits in day 3 anterior (blue), NMP (red) and NMP *T/Bra* KO (purple) samples. **(C)** Metaprofile illustrating mean accessibility within $\pm 2.5kb$ around spinal cord enhancer summits in day 5 anterior (blue), hindbrain (yellow), spinal cord (red) and spinal cord *T/Bra* KO (purple) samples. Figure adapted from [305].

In contrast, loss of CDX factors lead to a significant reduction in accessibilities of NMP and NMP/SC enhancers (NMP: $p < 2.22 \cdot 10^{-16}$, NMP-SC: $p < 1.3 \cdot 10^{-}16$, two-sided Wilcoxon rank; Fig. 4.15A, B, C). This confirmed the importance of these factors converting WNT signalling into chromatin accessibility. To understand the potential impact on gene expression as well as AP-axis regional identity, we profiled anterior and posterior marker genes. This confirmed that *Cdx* KO NPCs were unable to induce posterior *Hox* genes such *Hoxb9* and *Hoxc6* (Fig. 4.15D, E). However, the lack of CDX factors seemed to drive them more towards a hindbrain-like cell fate as indicated by expression of anterior *Hox* genes such as *Hoxb4* and *Hoxc4* (Fig. 4.15F, G).
Together this confirmed that *Cdx* KO cells were unable to provide access to NMP enhancers on day 3 and consequently failed to induce the necessary gene expression changes. Finally, these cells were unable to acquire a posterior NPC identity on day 5 and seemed to differentiate into hindbrain NPCs.

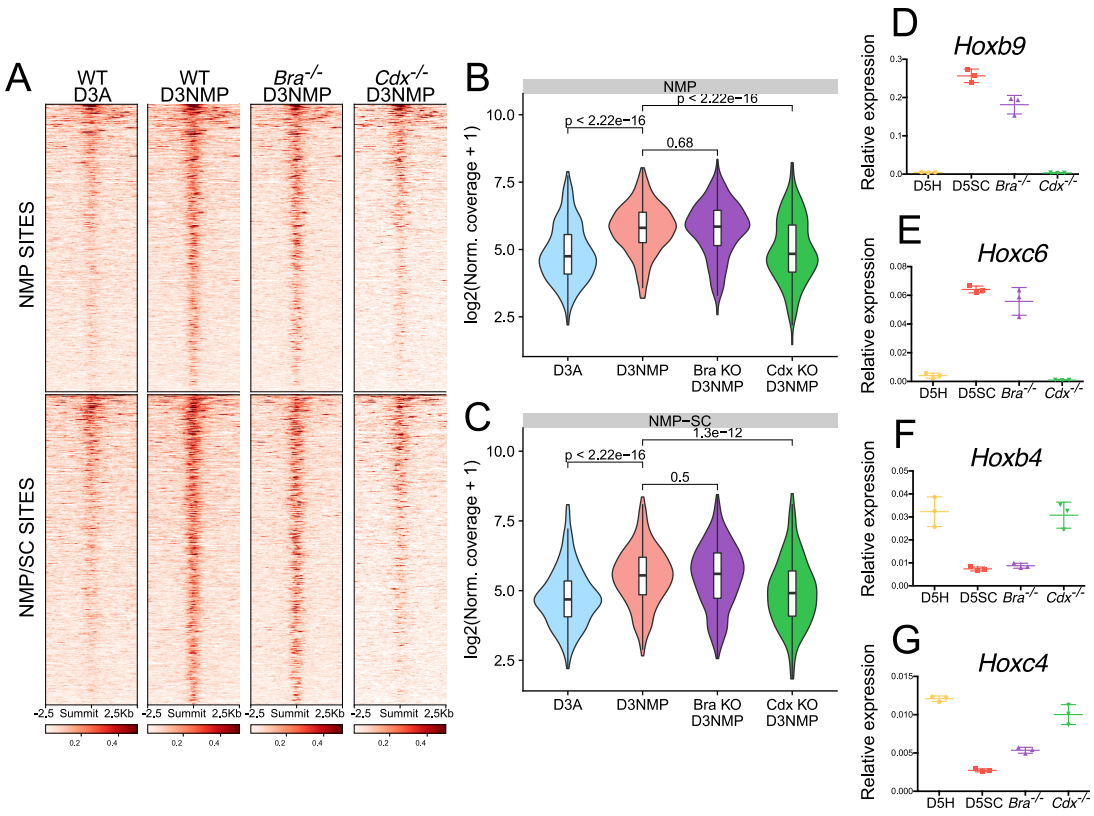


Fig. 4.15: CDX2 regulates NMP enhancer accessibility and spinal cord identity.
(A) Heatmaps displaying accessibility signal per base-pair $\pm 2.5kb$ around NMP and NMP-SC enhancer summits for day 3 anterior, NMP, NMP *T/Bra* KO and *Cdx* KO cells. **(B)** Quantification of chromatin accessibility in NMP enhancers for day 3 anterior, NMP, NMP *T/Bra* KO and *Cdx* KO cells, shown as violin plots (p-value: Wilcoxon rank sum test). **(C)** Quantification of chromatin accessibility in NMP-SC enhancers for day 3 anterior, NMP, NMP *T/Bra* KO and *Cdx* KO cells (p-value: Wilcoxon rank sum test). **(D-G)** qRT-PCR quantification of anterior (*Hoxb4* & *Hoxc4*) and posterior (*Hoxb9* & *Hoxc6*) marker genes in day 5 hindbrain, spinal cord, spinal cord *T/Bra* KO and *Cdx* KO cells. Expression is shown normalised to β-actin. Figure adapted from [305].

Following up our previous observation that *Cdx* KO ESC committed towards an anterior regional identity and differentiated to hindbrain NPCs, we wished to understand how chromatin accessibility changed in these cells. Fluorescence imaging of day 5 spinal cord *Cdx* and *T/Bra* KO revealed expression of the MNP marker *Olig2* in both, WT as well as *Cdx* KO cells (Fig. 4.16C). However, day 5 NPCs differentiated from *Cdx* KO ESCs showed expression of *Phox2b*, a known marker gene of visceral MNP, which supported the idea that these cells adopted a hindbrain identity, despite exposure to WNT signalling between day 2-3 (Fig. 4.16C). We used ATAC-seq from KO and WT conditions to define cell identity in the absence of *Cdx1/2/4*. Genome-wide comparison of spinal cord enhancer accessibility

revealed a decreased average accessibility in *Cdx* KO day 5 spinal cord, which was more comparable to hindbrain NPCs (Fig. 4.16C). On the other hand, hindbrain enhancers showed increased accessibility in day 5 *Cdx* KO cells compared with WT (Fig. 4.16D). Thus, cells lacking CDX differentiated into hindbrain NPCs and therefore an anterior identity. Moreover, CDX factors were necessary to remodel the chromatin landscape and prime these cells towards a posterior regional identity.

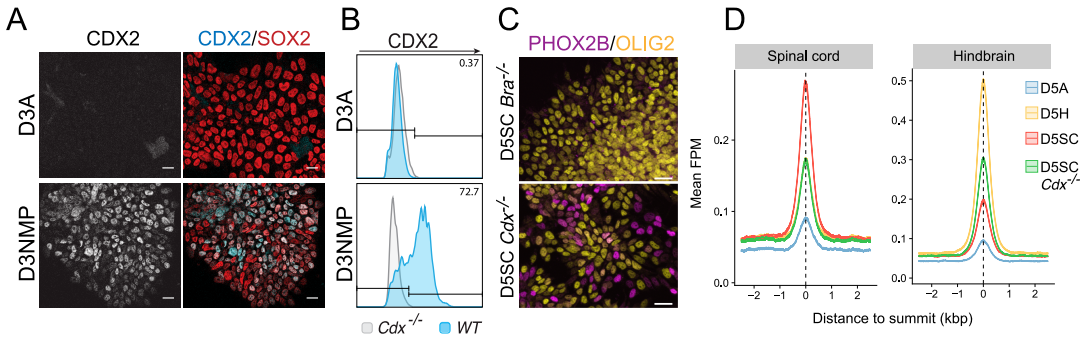


Fig. 4.16: *Cdx* KO cells differentiated towards hindbrain rather than spinal cord identity. **(A)** Fluorescence staining of CDX2 (cyan) SOX2 (red) in day 3 anterior and NMP cells (Scale $= 20\mu m$). **(B)** Flow cytometry quantification of CDX2 positive cells for WT (blue) and *Cdx* KO (grey) cells. **(C)** Staining of OLIG2 (yellow) and PHOX2B (magenta) in day 5 NPC differentiated from either T/Bra or Cdx KO cells by fluorescence imaging (Scale $= 20\mu m$). **D** Metaprofiles showing average accessibility of spinal cord and hindbrain enhancers in day 5 anterior, hindbrain, spinal cord and *Cdx* KO NPCs. Figure adapted from [305].

4.3.10 CDX2 inhibits hindbrain in favour of spinal cord identity

To understand the observed accessibility differences between WT and *Cdx* KO better, we mapped two different CDX2 ChIP-seq peak sets onto our SOM clusters (Fig. 4.17A). As expected, we observed the highest overlap of CDX2 binding sites with regulatory regions involved in the transition from NMP to SC. Surprisingly H/SC shared regulatory regions also overlapped with CDX2 ChIP-seq peaks. To exclude possible ChIP-seq artefacts, we compared signal enrichment for the different peak sets revealing comparable levels of enrichment for CDX2 binding sites independently of their SOM annotation (Fig. 4.18A).

Closer examination of CDX2 peaks overlapping H/SC regions led to the discovery of an enhancer element distal to *Phox2b* which might be directly involved in *Phox2b* repression (Fig. 4.17B). This suggested a possible mechanism of direct repression of hindbrain genes. Additionally, we observed CDX2 binding to a distal enhancer of *Mafb* upon induction of CDX2 in the hindbrain-like NPCs (Fig. 4.17C). Binding of CDX2 to hindbrain and shared

spinal cord enhancers suggests a dual role for the TF: on the one hand, it seems to be involved in the induction of posterior neuroectoderm and on the other site repression of an anterior identity [333, 315, 325, 360].

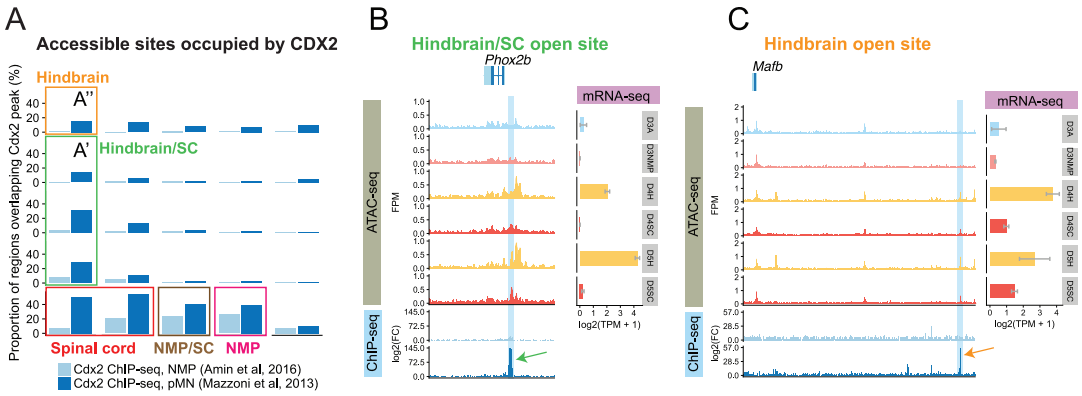


Fig. 4.17: CDX2 binding sites overlap with H-SC and hindbrain enhancers.
(A) Proportion of SOM clustered regions overlapping with CDX2 ChIP-seq peaks from NMPs (light blue) and motor neuron progenitors (dark blue) [315, 333]. **(B)** Example for a CDX2 binding shared H-SC enhancer proximal to *Phox2b*. Tracks show ATAC-seq accessibilities in day 3 A, day 3 NMP, day 4 H, day 4 SC, day 5 H and day 5 SC. *Right:* Barplot shows expression from *in vitro* RNA-seq (Error bar = SEM). *Bottom:* Plots show CDX2 ChIP-seq from NMP and MNP with induced CDX2 transgene. Colour shows experimental condition. **(C)** Example for a CDX2 binding hindbrain specific enhancer distal to *Mafb*. Figure adapted from [305].

To gain a better understanding of the genomic context in which CDX2 activates or represses expression, we performed motif enrichment analysis on Cdx2 peaks overlapping selected SOM clusters Fig. 4.18A). NMP, NMP-SC and spinal cord enhancers showed an expected enrichment of CDX and posterior HOX motifs (Fig. 4.18C). In contrast, hindbrain and H-SC sites were enriched for motifs of the SOX family (Fig. 4.18C). The only shared motif between hindbrain, spinal cord and H-SC enhancers was associated with retinoic acid receptor (RAR) which is in line with the induction of RA signalling in both differentiation protocols. However, RA signalling would be interpreted differently in these cells depending if they experienced WNT induction and already committed towards a posterior identity. In summary, these observations explained how CDX2 can influence interpretation of future signalling by priming the chromatin landscape for upregulation of posterior cell identities.
Next, we assessed which genes might be potential targets of CDX2, to address this we performed gene ontology enrichment analysis using GREAT [128]. These indicated differential functions of CDX2 bound enhancer regions depending on their accessibility (Fig. 4.18B). NMP and spinal cord enhancers seemed to be involved in WNT signalling regulation, whereas NMP-SC as well as spinal cord enhancers regulating genes responsible in establishing poste-

rior identity. On the other site, hindbrain and H-SC seem to be enriched for terms associated with anterior central nervous development such as "central nervous system development" or "regulation of neuron differentiation" (Fig. 4.18B).

To gain a better understanding of which potential target genes might be involved, we analysed gene expression datasets for *in vitro* hindbrain/spinal cord NPCs and CDX2 induced in hindbrain-like NPCs. This analysis revealed targets of CDX2 activation such as posterior *Hox* genes (*Hoxc8*, *Hoxa9*, *Hoxb9* and *Hoxc9*; Fig. 4.18D). Other CDX2 activated genes included the RA-producing enzyme *Aldh1a2* which was shown to be important for normal axis elongation [361]. Remarkably, the visceral MNP (hindbrain) marker *Phox2b* showed the strongest down-regulation after CDX2 induction in MNP (Fig. 4.18D). Together with the observation of CDX2 binding in close proximity to the *Phox2b* promoter, this might indicate how this strong downregulation was achieved and therefore *Phox2b* could be subject to direct CDX2 repression (Fig. 4.17B & Fig. 4.18D). In summary, these observations support a dual role for CDX2: on the one hand activating genes involved in spinal cord differentiation and on the other repressing genes important in establishing hindbrain fate.

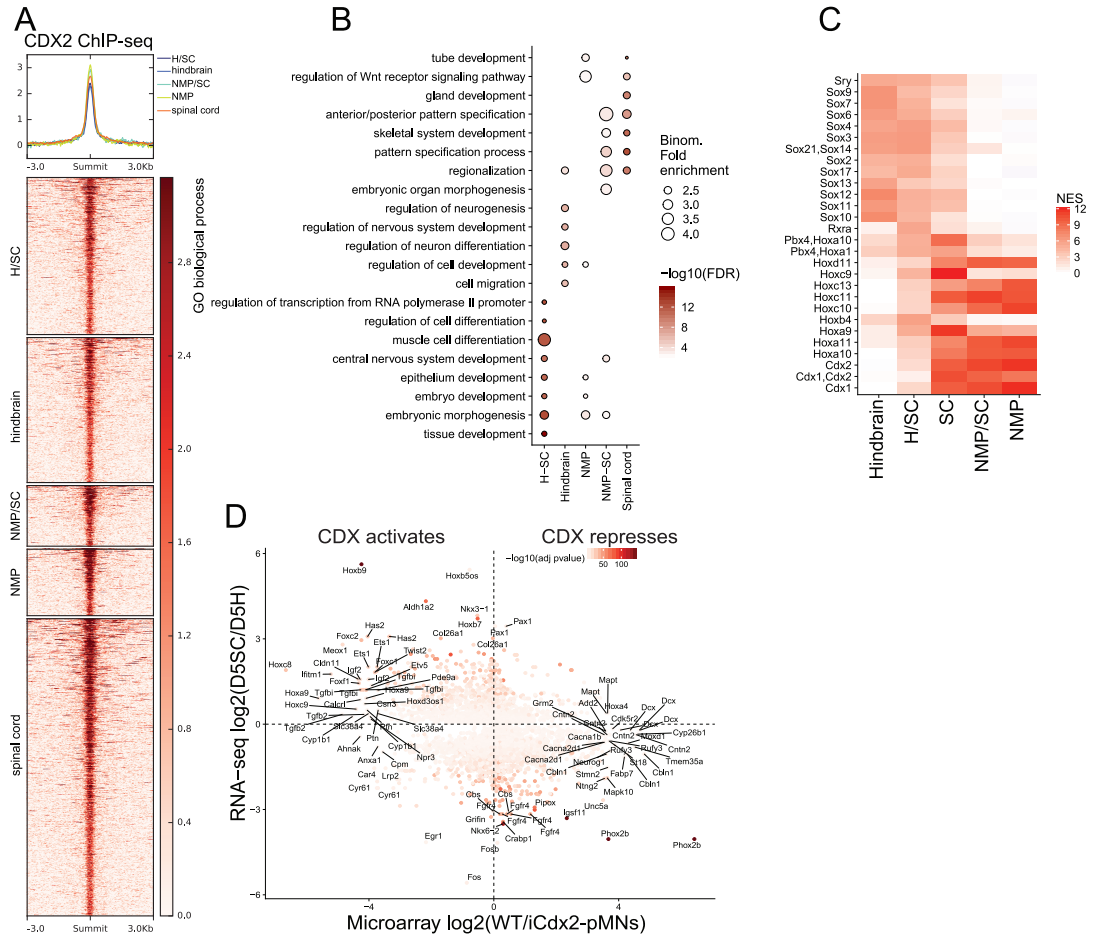


Fig. 4.18: CDX2 binds in different chromatin contexts to induce spinal cord and repress hindbrain identity.
(A) Heatmap showing normalised CDX2 ChIP-seq signal $\pm 3kb$ around the summit of CDX2 binding sites overlapping H-SC, hindbrain, NMP-SC, NMP and spinal cord enhancers. Metaprofile summarising CDX2 ChIP-seq signals is shown above. **(B)** Gene ontology enrichment of genes nearby to CDX2 binding sites overlapping with either H-SC, hindbrain, NMP-SC, NMP or spinal cord enhancers (p-value: Binomial test). **(C)** Motif enrichment within CDX2 binding sites overlapping with either H-SC, hindbrain, NMP-SC, NMP or spinal cord enhancers. Normalised enrichment scores (NES) were computed with iCisTarget. **(D)** Scatter plot comparing expression differences between spinal cord and hindbrain NPCs (y-axis) against induction of CDX2 in hindbrain-like NPCs (x-axis). Colour scale corresponds to -log10(FDR) derived from DESeq2 analysis of day5 SC and H RNA-seq. Figure adapted from [305].

4.3.11 CDX2, not WNT signalling, provides cells with the competence to generate spinal cord

CDX2 appears to be the crucial TF for activating genes involved in establishing spinal cord and repressing hindbrain identity. These observations led to the question of whether CDX2

induction alone is sufficient to differentiate cells towards a spinal cord identity, independently of developmental time. This would contrast the induction of WNT signalling whose effects on cell identity strongly depend on developmental time and context (Fig. 4.11 & Fig. 4.12). To answer these questions, we took advantage of our *in vitro* NPC differentiation protocol. If CDX2 is really the crucial TF operating downstream of WNT signalling it should be able to replace WNT induction at day 2. Therefore, we performed spinal cord NPC differentiation using a *Cdx2* inducible ESC cell line (iCDX2), harbouring a tamoxifen-inducible CDX2 transgene [362]. Similar to WT ESCs, iCDX2 ESCs generate vMNs expressing PHOX2B when exposed to hindbrain conditions. By contrast, the addition of WNT between day 2-3 suppresses the generation of these cells under the spinal cord protocol (Fig. 4.19A). Strikingly, we found that CDX2 induction between D2-3, was also sufficient to induce spinal cord (SCind; Fig. 4.19A). Both types of spinal cord NPCs showed expression of *Sox2* and lacked *Phox2b* (Fig. 4.19A, C). Moreover, gene expression profiling revealed similar levels of the motor neuron progenitor marker *Olig2* and successful induction of posterior *Hox* genes such as *Hoxc6* and *Hoxb9* (Fig. 4.19C). We performed ATAC-seq for SCind NPCs resulting in two replicates passing quality control with 28,426 consensus peaks (Fig. 4.4). This confirmed that CDX2 induction promote the same chromatin dynamics as observed in WT SC cells on a genome-wide scale. Day 5 SCind NPCs showed high accessibility in spinal cord and less accessibility in hindbrain enhancers (Fig. 4.19D). Accordingly, this recovered the same trend shown by WNT induced spinal cord NPCs compared with hindbrain NPCs. Taken together, our data confirmed that CDX2 alone was able to establish cells with a spinal cord identity. This raises the question of whether CDX2 activity was able to extend the period of competence that cells have to generate spinal cord.

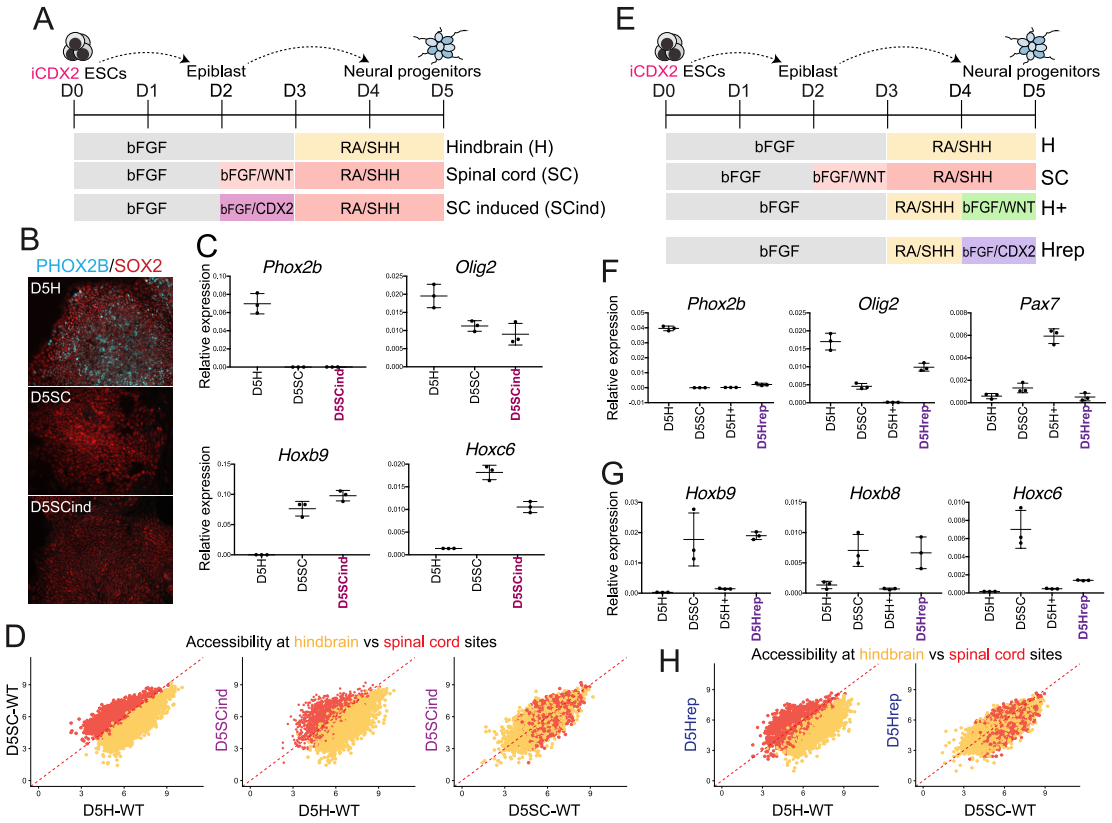


Fig. 4.19: CDX2 can replace WNT signalling to induce spinal cord and repress hindbrain identity.
(A) Schematic illustrating differentiation of spinal cord NPCs by induction of CDX2 after day 2 (SCind, purple). Differentiation of hindbrain and WNT based spinal cord NPCs shown. **(B)** Fluorescence staining of PHOX2B (magenta) and SOX2 (red) shown for day 5 hindbrain, spinal cord and SCind NPC. **(C)** qRT-PCR quantification of *Phox2b*, *Olig2*, *Hoxb9* and*Hoxc6* in hindbrain, spinal cord and SCind NPCs. **(D)** Scatter plot showing accessibility of hindbrain and spinal cord enhancers for day 5 hindbrain, spinal cord and SCind. **(E)** Schematic illustrating differentiation of hindbrain repressed NPCs by induction of *Cdx2* after day 4 (Hrep, light purple). Differentiation of hindbrain, spinal cord and hindbrain-plus NPCs shown. **(F)** qRT-PCR quantification of *Phox2b*, *Olig2*, *Pax7* for day 5 hindbrain, spinal cord, hindbrain-plus and hindbrain-repressed. Expression is shown normalised to β-actin. **(G)** qRT-PCR quantification of *Hoxb9*, *Hoxb8*, *Hoxc6* for day 5 spinal cord, hindbrain-plus and hindbrain-repressed. Expression is shown normalised to β-actin. **(H)** Scatter plot comparison of hindbrain (yellow) and spinal cord (red) enhancer accessibility for day 5 hindbrain, hindbrain repressed and spinal cord. Figure adapted from [305].

In order to test this hypothesis, we instead induced CDX2 in cells that had been differentiated under hindbrain conditions to day 4, a condition where the activation of WNT signalling will no longer posteriorise these cells (Fig. 4.18). Briefly, cells were differentiated into day 4 hindbrain NPC, followed by induction of CDX2. We referred to this condition as hindbrain repressed (Hrep; Fig. 4.19E). Initial gene expression analysis revealed that day 5 Hrep cells

maintained a neural identity, expressing markers such as *Olig2*, but lacked the expression of hindbrain marker *Phox2b* suggesting that NPCs lost their hindbrain identity (Fig. 4.19F). However, there was no observable expression of *Pax7* indicating that CDX2 induction was not comparable to WNT activation anymore at this time point. Additionally, cells that had been exposed to CDX2 expressed posterior *Hox* genes indicative of spinal cord fate, such as *Hoxb9* and *Hoxb8* (Fig. 4.19G). We performed ATAC-seq enabling genome-wide comparison of accessibility changes upon CDX2 induction in hindbrain NPC. Day 5 hindbrain repressed ATAC-seq samples passed our quality control and displayed 22,359 consensus peaks between replicates. This revealed that CDX2 induction in day 5 hindbrain NPCs induces chromatin changes that resemble spinal cord cell fate, losing accessibility in hindbrain enhancers but opening spinal cord enhancers (Fig. 4.19H).

In summary, timed induction of CDX2 was sufficient to establish a spinal cord identity in cells that would otherwise only transiently acquire the competence to make spinal cord NPCs. This was not dependent on WNT signalling, but rather, on CDX ensuring the accessibility of spinal cord enhancers.

4.4 Discussion

Here we used an *in vitro* model of neural induction combined with open chromatin profiling by ATAC-seq to challenge the "Activation Transformation" hypothesis and determine the exact sequence of events that generate NPCs with different regional identities along the AP-axis [308, 310]. The "Activation Transformation" hypothesis postulates that all NPCs initially acquire an anterior forebrain neural identity ("Activation"), before developing into either midbrain, hindbrain or spinal cord NPCs ("Transformation"; Fig. 4.1) [308]. Previous research supporting this hypothesis was mostly undertaken in amphibian or chicken embryos using embryology assays based on ex- or transplants [363, 364, 365]. However, an issue common to all of these studies is that the most posterior markers which were assayed correspond to hindbrain or anterior spinal cord identity. Findings of these studies were then extrapolated to the posterior spinal cord. Moreover, these experimental approaches lack the spatial and temporal resolution to thoroughly define cellular identity.

Over the last decade many methods based on high-throughput sequencing have been developed to assay the transcriptome (e.g. RNA-seq) and open-chromatin landscape (e.g. DNAse-seq, ATAC-seq) which can be used to more stringently define cellular identity [149, 82]. Here we utilized open chromatin profiling by ATAC-seq rather than gene expression profiling, as previous research had shown that regulatory regions were more predictive for cellular identity [167, 168, 169]. For example, Corces *et al* profiled chromatin accessibility and gene expression in different cell types during normal haematopoiesis from different donors [169]. In this study a ground truth of cell identity was determined by FACS sorting these cells based on surface markers before profiling. Cluster analysis using RNA-seq or ATAC-seq demonstrated that clustering using accessibility of distal regions (Cluster purity = 90.9%) outperforms clustering based on gene expression (Cluster purity = 77.6%).
In the case of neural development, it was shown that whilst factors such as SOX2 are broadly expressed throughout the developing nervous system, in fact different enhancer sets drive SOX2 expression in different neural cells [321, 322, 323]. Hence, chromatin accessibility of enhancer elements seems to be more predictive for cellular identity than gene expression.

Initially, we confirmed that our *in vitro* differentiation model undergoes similar chromatin remodelling events as corresponding *in vivo* cells. This was demonstrated by loss of accessibility at a well-known pluripotency enhancer distal to *Pou5f1* and gain of accessibility upon differentiation in known neural enhancers regulating *Olig2* (Fig. 4.3) [216, 337]. These

observations are in line with previous published single cell gene expression profiling studies, thus confirming the concordance of *in vitro* generated NPCs with their *in vivo* counterparts [318, 311].

In this study, we used a self-organising map (SOM) to cluster time-series chromatin accessibility data. Previous studies on gene expression data showed high accuracy for SOM based clustering [366, 367]. Moreover, we compared SOM and k-means clustering and found the SOM clustering to be more biologically meaningful. One possible reason for this is that a SOM-based clustering gives a topological map of the clustered data in which neighbouring clusters share similar trends. Given that our data is a developmental time course, we expect such an approach to return a much more informative clustering than k-means or hierarchical clustering.
The unsupervised SOM clustering of dynamic accessible regions during this differentiation identified nine distinct biologically meaningful clusters that resolved the temporal dynamics. Three clusters corresponded to regulatory elements specific for NPCs with different AP-axis regional identities: anterior, hindbrain and posterior spinal cord. Further analysis uncovered key TFs that shape the chromatin landscape of these cells. For example, spinal cord enhancers showed enrichment for CDX2 as well as HOXC9 motifs and ChIP-seq peaks. Both factors have been described as important TFs in posterior spinal cord development. For example, loss of CDX2 during mouse embryonic developments results in impaired axis elongation [315, 311]. In contrast, hindbrain and spinal cord enhancers are enriched for SOX2 binding sites, in line with the ubiquitous expression of Sox factors along the AP-axis [339]. Moreover, this demonstrates how a broadly expressed TF is involved in mediating different functions and promoting different cell identities depending in the sequence context its binding [322]. Together, we identified enhancer elements as well as their corresponding master regulators which are responsible for establishing AP-axis regional identity.

Furthermore, we observed a set of transient regulatory regions that opened up upon exposure to WNT signalling on day 3 of our *in vitro* NPC differentiation protocol. These day 3 cells correspond to a previously described bipotent NMP state which is highlighted by the expression of *Cdx*, *T/Bra* and *Sox2* and gives rise to posterior spinal cord NPC and paraxial mesoderm [310, 309]. Further analysis of *Cdx1,2,4* and *T/Bra* KO ATAC-seq highlighted the importance of the NMP state and CDX factors for generating NPC with a spinal cord identity. Epiblast cells lacking CDX are not able to integrate WNT signalling, failing to induce a posterior identity and instead differentiate towards hindbrain-like NPCs. This is consistent with

previous work showing that posterior development is impaired upon *Cdx1,2,4* KO or absence of WNT signalling, meaning that the NMP state is essential for normal tail development, especially for generating the posterior spinal cord [314, 315, 325]. Therefore, pluripotent epiblast cells need to commit first towards a posterior identity ("primary regionalization") before these cells can acquire neural identity (Fig. 4.20A).

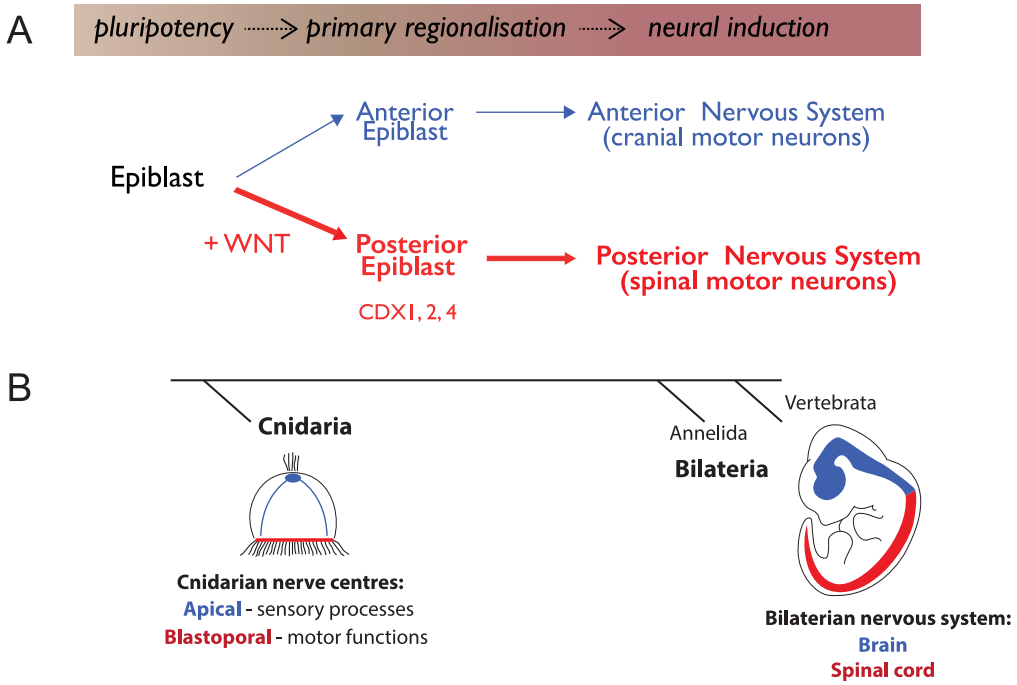


Fig. 4.20: Proposed model of axial allocation and neural induction during early neurogenesis. **(A)** Illustrated summary of the proposed neurogenesis model. Initially pluripotent epiblast cells first commit either to an anterior or posterior regional identity depending on presence or absence of WNT signalling. This leads to anterior epiblast cells which give rise to neuroectoderm forming the anterior nervous system (blue). In the presence of WNT signalling, epiblast cells differentiate to NMPs by induction of CDX factors. These cells can give rise to posterior neuroectoderm which forms our spinal cord. **(B)** Evolutionary perspective on the proposed model of axial allocation and neural induction during early neurogenesis. The model describes the dual origin of nervous system hypothesis. According to this hypothesis was the posterior nervous system once separated from the anterior one which is still observable in Cnidaria. In these animals there is still an Apical (blue) system for sensory processes and Blastoporal (red) system for motor functions. However, in bilateria these nervous systems are connected as the brain (blue) and spinal cord (red). Figure adapted from [305].

We identified CDX2 as master regulator of posterior identity. Using a tamoxifen inducible CDX2 mESC cell line, we showed that WNT signalling on day 3 can be replaced by timed

induction of CDX2 to induce the NMP state. Strikingly, induction of WNT signalling on day 5 failed to induce posterior identity in hindbrain NPCs. However, induction of CDX2 in day 5 hindbrain NPCs leads to develop a posterior spinal cord identity. Hence, we showed that CDX2 alone and not WNT signalling is the crucial factor for establishing a posterior neural identity. Investigation into potential underlying mechanisms of these observations revealed a dual role of CDX2 in establishing spinal cord identity. On the one hand, CDX2 activates genes involved in establishing a posterior neural identity such as *Hoxc9*. On the other hand, CDX2 represses hindbrain identity genes such as *Phox2b* in favour of a posterior identity by binding to shared hindbrain/spinal cord enhancer sites.

The three CDX genes (*Cdx1*, *Cdx2* and *Cdx4*) in vertebrates today were created over 500 million years ago after several whole genome duplications took place [368]. These CDX genes originated from a single CDX gene in vertebrate ancestors and can still be found in cephalochordates such as *Branchiostoma floridae* [369]. Moreover, homologous of the mammalian CDX genes can be found along the tree of life, for example in nematodes such as *Caenorhabditis elegans* (*pal-1*), in fruit flies such as *Drosophila* (*cad*) and two copies in fish such as zebrafish (*cdx1* and*cdx4*) [370, 369, 371]. Previous studies showed that even the WNT induced CDX gene regulatory network is evolutionarily conserved among animals [372, 353]. Thus, the distinct mechanisms leading to the development of an anterior and posterior nervous system in vertebrates might be a remnant of processes which generated an apical and blastoporal system in pre-bilaterians. Our observations provide further support for the hypothesis that the central nervous system in bilaterians has evolved as a merger of two initially distinct nervous systems that originated at opposite poles in pre-bilaterian ancestors [373] (Fig. 4.20B).

In summary, our findings contradict the "Activation Transformation" hypothesis and support the validity of Mangold's model/Dual origin of NPCs: that distinct mechanisms are necessary to generate different regional identities along the AP-axis.

5

Conclusions and future directions

In this PhD thesis, I investigated how regulatory potential and cell type specificity is encoded in the underlying DNA sequence of transcriptional enhancers. To this end, enhancers in the context of two pluripotent ESC states and during early neural development were examined using high-throughput sequencing assays such as ATAC-seq and ChIP STARR-seq.
In Chapter 2, I asked whether all NANOG binding sites observed in ChIP-seq data correspond to *bona fide* enhancer elements and what DNA sequence features encode regulatory activity in NANOG-bound enhancers. NANOG binding sites were systematically tested for their ability to drive transcription from a minimal promoter using ChIP STARR-seq. This allowed for a thorough definition of NANOG binding sites with enhancer activity and revealed inactive binding sites which were highly bound by NANOG but did not display regulatory activity in the cellular state under examination. I characterised these two classes of NANOG binding sites and built machine learning models to uncover predictors of enhancer activity. This revealed ESRRB as important regulator of enhancer activity which I validated using publicly available H3K27ac ChIP-seq data measured in mESC harbouring an inducible *Esrrb* KO.
In Chapter 3, we examined how differential enhancer activity between naive (2i) and primed (serum) mESCs is encoded in the DNA sequences underlying NANOG binding sites. Enhancer activity of NANOG binding sites in 2i and serum conditions were measured using ChIP STARR-seq. Identified 2i- and serum-specific enhancer elements were examined, revealing chromatin accessibility changes as a major predictor for differential enhancer activity. Predictive modelling of differential enhancer activities revealed the TFs TCF3, TCF4 and ESRRB as potential regulators of 2i-specific activity and ZIC3 for serum enhancers. Together, Chapters 2 and 3 presented an analysis framework to identify TF-bound enhancers and uncover predictive features that may explain how they switch between regulatory and non-regulatory states.

Finally, in Chapter 4, in collaboration with James Briscoe's laboratory, I assessed whether it is possible to use information about enhancer accessibilities to define precisely when cells commit to specific neural identities. Application of ATAC-seq to an *in vitro* model system of early neural development uncovered enhancer elements specific to the mouse forebrain, hindbrain and spinal cord NPC identity and identified TFs associated with cell type specific enhancer accessibility. By perturbing TFs and signalling factors, I showed that that cells first commit to a regional (posterior) identity before then specifying the cell type (neural) identity to generate spinal cord NPCs. My analysis showed that this differentiation goes via the bipotent NMP state with CDX2 as master regulator of posterior identity by repressing hindbrain and activating spinal cord genes. The work in this chapter confirmed the "Dual origin of neural progenitor" hypothesis and rejected the more commonly accepted "Activation-Transformation" hypothesis of neural development.

Overall, this thesis has made extensive use of ATAC-seq and ChIP STARR-seq data to study enhancer elements in different biological contexts. Both of these techniques come with their own technical obstacles and limitations to define *bona fide* transcriptional enhancers. The two biological contexts each provided its own challenges and setbacks for these techniques. In the following sections I discuss these issues, suggest potential solutions and finally discuss future directions for this area of research.

5.1 Improving ChIP STARR-seq to achieve high resolution readouts

As mentioned in Chapters 2 and 3 our implementation of the ChIP STARR-seq assay comes with two experimental obstacles. Firstly, it was recently shown that the bacterial origin of replication (ORI) found in original version of the mammalian STARR-seq screening vector functions as an alternative core-promoter; this competes with the minimal core-promoter from which STARR-seq measurements are supposed to be made, leading to a signal reduction [176]. In future, this is resolved by using the ORI as the minimal core-promoter. Using such a vector will improve signal yield in future STARR-seq screens for transcriptional enhancers in mammalian cells.
Secondly, the lack of UMIs to distinguish PCR duplicates from independently transcribed fragments reduces the recoverable STARR-seq signal and the potential assay resolution [219, 184]. In this thesis, I had to average the STARR-seq signal over an entire binding

site and therefore quantified how many fragments under a particular peak were able to self-transcribe. Introducing UMIs would enable us to quantify each cloned fragment as a separate measurement since we would be able to quantify self-transcription of each observed fragment. This in turn would increase the resolution of the assay, by tilling each binding peak with STARR-seq fragments, thus allowing the fine-mapping of individual "regulatory driver elements" [256, 179]. In our case, fragment tilling of TF binding peaks could eventually provide support for the base-pair resolution importance score from DeepLIFT. For example, fragments overlapping an ESRRB motif with high DeepLIFT importance should show higher expression than fragments within the same peak that do not overlap this motif.

A second advantage will be the increase in measurements: instead of thousands of data points one might have up to hundreds of thousands. This would allow accurate fitting of larger and more complex models such as recurrent neural networks (for instance, long short-term memory models) or enable models for more complex tasks such as regression of enhancer activity instead of binary prediction [210]. A recent study presented evidence that it might be possible to gain a better understanding of the motif grammar encoded in DNA sequence by combining high resolution measurements of TF-binding measurements with predictive modelling using CNNs [374]. However, they also relied on several 100,000 binding measurements to achieve this.

5.2 Validation of ChIP STARR-seq and predicted regulators of enhancer activity

For Chapter 2, I plan to validate the NANOG binding site classification by selecting 10 active enhancers and 10 inactive TFBS, and measuring their enhancer activities using a dual luciferase reporter assay [375, 180]. A limitation of course, is that this will not assess enhancer activities on endogenous gene expression. For this, it would be necessary to modify binding sites using CRISPR/Cas9 or selectively silence/deactivate them with dCas9-effector fusions such as dCas9-KRAB or dCas9-LSD1 [224, 376, 60, 377]. It will be interesting to see if deletion of inactive binding sites impacts gene expression in ESCs at all and if removal of these sites for example, the known *Hoxa1* enhancer is sufficient to impair differentiation [222]. Even such an experiment might not be definitive because both of these endpoints might be rescued by so called shadow enhancer elements standing in for deleted or silenced enhancers [42].

I also plan to validate my regulator predictions based on the importance scores using dual

luciferase assays. I will determine the most important sequence motifs in an enhancer using the DeepLIFT base-pair resolution importance scores, shuffle these sequences and measure their episomal enhancer activity [216]. In the case of active enhancers, I will alter candidate sequences targeting the discovered ESRRB motifs in a way that minimises the predicted probability of being an enhancer by the CNN. For inactive binding sites I will search for unimportant sections of sequence and introduce ESRRB motifs so as to maximise the predicted probability to be an active enhancer. This experimental strategy should confirm the role of the discovered ESRRB motifs in mediating enhancer activity. However, sequence motifs of the nuclear receptor family such as ESRRB show a high degree of similarity between each other, such as ESRRB and NR0B1 [199, 63]. Hence, deleting or inserting ESRRB motifs from candidate sites might not conclusively show that ESRRB establishes enhancer activity. To establish such causality, I will also measure the set of altered active enhancer and inactive TFBS sequences using a transgenic ESC line in which we are able to degrade ESRRB by a light inducible system [378].

Another potential approach to validate the base-pair importance score from DeepLIFT as well as the importance of the discovered ESRRB motifs would be to perform a tiling based CRISPR/Cas9 screen of selected candidate regions. Two recently published studies described screens to map regulatory regions using different endpoint readouts [379, 380].

The above described experimental procedures could also be applied to validate our predictions in Chapter 3, with two alterations: using a base-pair importance score based on the gkm-SVM from GkmExplain and measuring enhancer activities in 2i and serum conditions [302].

5.3 Definition of *bona fide* transcriptional enhancers from ATAC-seq

In Chapter 4, we used open-chromatin profiling by ATAC-seq to identify regulatory elements via chromatin accessibility. However, these distal regulatory regions are not necessarily enhancers and might correspond to insulators or silencers [381, 382, 82]. A powerful new approach called ATAC-STARR-seq was presented by Wang *et al*, which combines enrichment of open-chromatin regions via Tn5 transposase with enhancer screening via STARR-seq [179]. The authors did not comment on how many distal peaks show self-transcription in their STARR-seq assay. Application of this method to our *in vitro* model system would uncover what proportion of the SOM regions actually are *bona fide* transcription al enhancers potentially impacting gene expression. Moreover, this might help to decipher how CDX2 is

represses hindbrain genes such as *Phox2b* and *Mafb* (Fig. 4.17). It would be interesting to see how the shared hindbrain/spinal cord shared regulatory regions function when exposed to the cellular conditions specific to the hindbrain or spinal cord. Ultimately, a combination of such measurements with *in vitro* models of developmental processes would allow for more fine-grained classification of regulatory regions and uncover the potential contribution of enhancers, silencers and insulators during development.

5.4 Resolving cellular heterogeneity using single cell ATAC-seq

Another point regarding Chapter 4, concerns our *in vitro* system of early neural development: as a cell culture system, there might be substantial cell-to-cell heterogeneity in the dish [383, 384, 385]. Different subpopulations in cell culture could arise through uneven dosage of treatment across a dish or differences in timing or length of exposure to certain signals. Population level methods such as ChIP-seq or ATAC-seq will result in averaged measurements across potentially heterogenous populations. This may result in higher background levels or produce a convoluted signal showing chromatin accessibility at various regions that are actually specific to different subpopulations. Such issues might apply to our study as well. For example, one could speculate that our observed shared hindbrain/spinal cord regulatory regions might actually originate from cells which did not commit on day 3 to a posterior identity and therefore generated hindbrain NPCs upon treatment with RA and SHH. This would result in a mixed population of hindbrain and spinal cord NPCs. However, this is only speculation and it is interesting that on day 3 after the induction of WNT signalling, cells commit to a bipotent NMP cell state which gives rise to spinal cord NPCs and paraxial mesoderm [310, 311]. Commitment of NMPs to paraxial mesoderm is regulated by continuous WNT signalling which might have happened in the dish leading to a few cells committing towards a mesodermal rather than spinal cord identity. Thus, it is also possible for cell-to-cell heterogeneity to arise from normal developmental decisions. Recent advances in single cell sequencing had made it possible to perform ATAC-seq in single cells [133]. Applying single cell ATAC-seq (scATAC-seq) to our *in vitro* system will have advantages: 1) it will resolve the issue of potential cellular heterogeneity originating either from the culture system or via developmental processes; and 2) it will allow us to investigate these developmental processes at higher temporal resolution; for instance, what precise order of accessibility gains is necessary for spinal cord identity? The possibility of gaining a better understanding of

developmental processes using scATAC-seq was exemplified by a recent study investigating myoblast differentiation with an *in vitro* system [384].

5.5 Concluding remarks

There are many different definitions for transcriptional enhancer elements based on various different computational methods or sequencing assays but none of them is yet definite [161, 170, 163]. Therefore, it is and will be still necessary to further develop techniques to thoroughly map transcriptional enhancers and their associated target genes. Only one thing is certain in the future, that the right choice of assay and associated enhancer definition will still strongly depend on the research question asked.

Bibliography

[1] Matthew Cobb. “60 years ago, Francis Crick changed the logic of biology”. In: *PLOS Biology* 15.9 (2017), e2003243.

[2] François Jacob and Jacques Monod. “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356.

[3] James T. Kadonaga. “Perspectives on the RNA polymerase II core promoter”. In: *Wiley Interdisciplinary Reviews: Developmental Biology* 1.1 (2012), pp. 40–51.

[4] Ann E. Rougvie and John T. Lis. “The RNA polymerase II molecule at the 5’ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged”. In: *Cell* 54.6 (1988), pp. 795–804.

[5] E. B. Rasmussen and J. T. Lis. “In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 90.17 (1993), pp. 7923–7927.

[6] Nadine Czudnochowski, Christian A. Boesken, and Matthias Geyer. “Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition”. In: *Nature Communications* 3.1 (2012), pp. 1–12.

[7] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. “Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters”. In: *Science* 322.5909 (2008), pp. 1845–1848.

[8] Karen Adelman and John T Lis. “Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.” In: *Nature Reviews Genetics* 13.10 (2012), pp. 720–31.

[9] N F Marshall and D H Price. “Control of formation of two distinct classes of RNA polymerase II elongation complexes.” In: *Molecular and Cellular Biology* 12.5 (1992), pp. 2078–2090.

[10] B. Matija Peterlin and David H. Price. *Controlling the Elongation Phase of Transcription with P-TEFb*. 2006.

[11] Ivan Gusarov and Evgeny Nudler. “The mechanism of intrinsic transcription termination”. In: *Molecular Cell* 3.4 (1999), pp. 495–504.

[12] N. J. Proudfoot. “Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut”. In: *Science* 352.6291 (2016), aad9926–aad9926.

[13] Boris Lenhard, Albin Sandelin, and Piero Carninci. "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." In: *Nature Reviews Genetics* 13.4 (2012), pp. 233–45.

[14] Benjamin S. Scruggs, Daniel A. Gilchrist, Sergei Nechaev, Ginger W. Muse, Adam Burkholder, David C. Fargo, and Karen Adelman. "Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin". In: *Molecular Cell* 58.6 (2015), pp. 1101–1112.

[15] Amy C. Seila, J. Mauro Calabrese, Stuart S. Levine, Gene W. Yeo, Peter B. Rahl, Ryan A. Flynn, Richard A. Young, and Phillip A. Sharp. "Divergent transcription from active promoters". In: *Science* 322.5909 (2008), pp. 1849–1851.

[16] Leighton J. Core, Andre L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers". In: *Nature Genetics* 46.12 (2014), pp. 1311–1320.

[17] Alla A. Sigova, Alan C. Mullen, Benoit Molinie, Sumeet Gupta, David A. Orlando, Matthew G. Guenther, Albert E. Almada, Charles Lin, Phillip A. Sharp, Cosmas C. Giallourakis, and Richard A. Young. "Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 110.8 (2013), pp. 2876–2881.

[18] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage". In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15776–15781.

[19] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A.M. Semple, Martin S. Taylor, Paer G. Engstroem, Martin C. Frith, Alistair R.R. Forrest, Wynand B. Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M. Grimmond, Christine A. Wells, Valerio Orlando, Claes Wahlestedt, Edison T. Liu, Matthias Harbers, Jun Kawai, Vladimir B. Bajic, David A. Hume, and Yoshihide Hayashizaki. "Genome-wide analysis of mammalian promoter architecture and evolution". In: *Nature Genetics* 38.6 (2006), pp. 626–635.

[20] Vanja Haberle and Alexander Stark. "Eukaryotic core promoters and the functional basis of transcription initiation". In: *Nature Reviews Molecular Cell Biology* (2018), p. 1.

[21] R. P. Lifton, M. L. Goldberg, R. W. Karp, and D. S. Hogness. "The organization of the histone genes in Drosophila melanogaster: Functional and evolutionary implications". In: *Cold Spring Harbor Symposia on Quantitative Biology* 42.2 (1977), pp. 1047–1051.

[22] Georgia A. Patikoglou, Joseph L. Kim, Liping Sun, Sang Hwa Yang, Thomas Kodadek, and Stephen K. Burley. "TATA element recognition by the TATA box-binding protein has been conserved throughout evolution". In: *Genes and Development* 13.24 (1999), pp. 3217–3230.

[23] Long Vo Ngoc, California Jack Cassidy, Cassidy Yunjing Huang, Sascha H.C. Duttke, and James T. Kadonaga. "The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters". In: *Genes and Development* 31.1 (2017), pp. 6–11.

[24] Thierry Lagrange, Achillefs N. Kapanidis, Hong Tang, Danny Reinberg, and Richard H. Ebright. "New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB". In: *Genes and Development* 12.1 (1998), pp. 34–44.

[25] Wensheng Deng and Stefan G.E. Roberts. "A core promoter element downstream of the TATA box that is recognized by TFIIB". In: *Genes and Development* 19.20 (2005), pp. 2418–2423.

[26] Dong-Hoon Lee, Naum Gershenzon, Malavika Gupta, Ilya P. Ioshikhes, Danny Reinberg, and Brian A. Lewis. "Functional Characterization of Core Promoter Elements: the Downstream Core Element Is Recognized by TAF1". In: *Molecular and Cellular Biology* 25.21 (2005), pp. 9674–9686.

[27] Robert K. Louder, Yuan He, Jose Ramon Lopez-Blanco, Jie Fang, Pablo Chacon, and Eva Nogales. "Structure of promoter-bound TFIID and model of human pre-initiation complex assembly". In: *Nature* 531.7596 (2016), pp. 604–609.

[28] M. Gardiner-Garden and M. Frommer. "CpG Islands in vertebrate genomes". In: *Journal of Molecular Biology* 196.2 (1987), pp. 261–282.

[29] Vanja Haberle, Nan Li, Yavor Hadzhiev, Charles Plessy, Christopher Previti, Chirag Nepal, Jochen Gehrig, Xianjun Dong, Altuna Akalin, Ana Maria Suzuki, Wilfred F. J. van IJcken, Olivier Armant, Marco Ferg, Uwe Straehle, Piero Carninci, Ferenc Mueller, and Boris Lenhard. "Two independent transcription initiation codes overlap on vertebrate core promoters". In: *Nature* 507.7492 (2014), pp. 381–385.

[30] Riu Yamashita, Yutaka Suzuki, Sumio Sugano, and Kenta Nakai. "Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity". In: *Gene* 350.2 (2005), pp. 129–136.

[31] Elizabeth A. Rach, Deborah R. Winter, Ashlee M. Benjamin, David L. Corcoran, Ting Ni, Jun Zhu, and Uwe Ohler. "Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level". In: *PLoS Genetics* 7.1 (2011). Ed. by Jason D. Lieb, e1001274.

[32] Julian Banerji, Sandro Rusconi, and Walter Schaffner. "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences". In: *Cell* 27.2 PART 1 (1981), pp. 299–308.

[33] J Banerji, L Olson, and W Schaffner. "A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes." In: *Cell* 33.3 (1983), pp. 729–40.

[34] Franziska Reiter, Sebastian Wienerroither, and Alexander Stark. “Combinatorial function of transcription factors and cofactors”. In: *Current Opinion in Genetics & Development* 43 (2017), pp. 73–81.

[35] Tae Kyung Kim, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. “Widespread transcription at neuronal activity-regulated enhancers”. In: *Nature* 465.7295 (2010), pp. 182–187.

[36] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David A Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. “An atlas of active enhancers across human cell types and tissues.” In: *Nature* 507.7493 (2014), pp. 455–461.

[37] Jen Hao Cheng, David Zhi Chao Pan, Zing Tsung Yeh Tsai, and Huai Kuang Tsai. “Genome-wide analysis of enhancer RNA in gene regulation across 12 mouse tissues”. In: *Scientific Reports* 5.1 (2015), pp. 1–9.

[38] Michael T.Y. Lam, Wenbo Li, Michael G. Rosenfeld, and Christopher K. Glass. “Enhancer RNAs and regulated transcriptional programs”. In: *Trends in Biochemical Sciences* 39.4 (2014), pp. 170–182.

[39] A. A. Sigova, B. J. Abraham, X. Ji, B. Molinie, N. M. Hannett, Y. E. Guo, M. Jangi, C. C. Giallourakis, P. A. Sharp, and R. A. Young. “Transcription factor trapping by RNA in gene regulatory elements”. In: *Science* 350.6263 (2015), pp. 978–81.

[40] Kambiz Mousavi, Hossein Zare, Stefania Dell’Orso, Lars Grontved, Gustavo Gutierrez-Cruz, Assia Derfoul, Gordon L. Hager, and Vittorio Sartorelli. “ERNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci”. In: *Molecular Cell* 51.5 (2013), pp. 606–617.

[41] Denes Hnisz, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. “A Phase Separation Model for Transcriptional Control”. In: *Cell* 169.1 (2017), pp. 13–23.

[42] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. “Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution”. In: *Cell* 167.5 (2016), pp. 1170–1187.

[43] D Thanos and T Maniatis. “Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome.” In: *Cell* 83.7 (1995), pp. 1091–100.

[44] David N. Arnosti and Meghana M. Kulkarni. “Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?” In: *Journal of Cellular Biochemistry* 94.5 (2005), pp. 890–898.

[45] Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, E Hilary Gustafson, Ewan Birney, and Eileen E M Furlong. “A transcription factor collective defines cardiac cell fate and reflects lineage history.” In: *Cell* 148.3 (2012), pp. 473–86.

[46] Muhammad A. Zabidi and Alexander Stark. “Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors”. In: *Trends in Genetics* 32.12 (2016), pp. 801–814.

[47] Muhammad A. Zabidi, Cosmas D. Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. “Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation”. In: *Nature* 518.7540 (2014), pp. 556–9.

[48] Cosmas D Arnold, Muhammad A Zabidi, Michaela Pagani, Martina Rath, Katharina Schernhuber, Tomáš Kazmar, and Alexander Stark. “Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution”. In: *Nature Biotechnology* 35 (2016), pp. 136–144.

[49] Vanja Haberle, Cosmas D. Arnold, Michaela Pagani, Martina Rath, Katharina Schernhuber, and Alexander Stark. “Transcriptional cofactors display specificity for distinct types of core promoters”. In: *Nature* (2019), p. 1.

[50] Jonathan R. Chubb, Tatjana Trcek, Shailesh M. Shenoy, and Robert H. Singer. “Transcriptional Pulsing of a Developmental Gene”. In: *Current Biology* 16.10 (2006), pp. 1018–1025.

[51] Gil Hornung, Raz Bar-Ziv, Dalia Rosin, Nobuhiko Tokuriki, Dan S. Tawfik, Moshe Oren, and Naama Barkai. “Noise-mean relationship in mutated promoters”. In: *Genome Research* 22.12 (2012), pp. 2409–2417.

[52] Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean Marc Victor, Marie Cecile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, Xavier Darzacq, Christophe Zimmer, Eugenia Basyuk, and Edouard Bertrand. “A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting”. In: *Nature Communications* 7.1 (2016), pp. 1–14.

[53] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. “Genomic encoding of transcriptional burst kinetics”. In: *Nature* (2019), p. 1.

[54] Caroline R. Bartman, Sarah C. Hsu, Chris C.-S. Hsiung, Arjun Raj, and Gerd A. Blobel. “Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping”. In: *Molecular Cell* 62.2 (2016), pp. 237–247.

[55] Takashi Fukaya, Bomyi Lim, and Michael Levine. “Enhancer Control of Transcriptional Bursting.” In: *Cell* 166.2 (2016), pp. 358–368.

[56] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong Ihn Lee, and Richard A Young. “Master transcription factors and mediator establish super-enhancers at key cell identity genes.” In: *Cell* 153.2 (2013), pp. 307–19.

[57] Denes Hnisz, Brian J Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A Sigova, Heather A Hoke, and Richard A Young. "Super-enhancers in the control of cell identity and disease." In: *Cell* 155.4 (2013), pp. 934–47.

[58] Violaine Saint-André, Alexander J Federation, Charles Y Lin, Brian J Abraham, Jessica Reddy, Tong Ihn Lee, James E Bradner, and Richard A Young. "Models of human core transcriptional regulatory circuitries". In: *Genome Research* 26.3 (2016), pp. 385–396.

[59] Noah Dukler, Brad Gulko, Yi-Fei Huang, and Adam Siepel. "Is a super-enhancer greater than the sum of its parts?" In: *Nature Genetics* 49.1 (2016), pp. 2–3.

[60] Sakthi D. Moorthy, Scott Davidson, Virlana M. Shchuka, Gurdeep Singh, Nakisa Malek-Gilani, Lida Langroudi, Alexandre Martchenko, Vincent So, Neil N. Macpherson, and Jennifer A. Mitchell. "Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes". In: *Genome Research* 27.2 (2017), pp. 246–258.

[61] Carsten Bahr, Lisa von Paleske, Veli V. Uslu, Silvia Remeseiro, Naoya Takayama, Stanley W. Ng, Alex Murison, Katja Langenfeld, Massimo Petretich, Roberta Scognamiglio, Petra Zeisberger, Amelie S. Benk, Ido Amit, Peter W. Zandstra, Mathieu Lupien, John E. Dick, Andreas Trumpp, and François Spitz. "A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies". In: *Nature* 553.7689 (2018), pp. 515–520.

[62] Mark Ptashne. "Specific binding of the λ phage repressor to λ DNA". In: *Nature* 214.5085 (1967), pp. 232–234.

[63] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. "The Human Transcription Factors". In: *Cell* 172.4 (2018), pp. 650–665.

[64] Matthew T. Weirauch and T. R. Hughes. "A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution". In: *Sub-Cellular Biochemistry* 52 (2014), pp. 25–73.

[65] Thomas Edlund, Michael D. Walker, Philip J. Barr, and William J. Rutter. "Cell-specific expression of the rat insulin gene: Evidence for role of two distinct 5' flanking elements". In: *Science* 230.4728 (1985), pp. 912–916.

[66] Seth Frietze and Peggy J. Farnham. "Transcription factor effector domains". In: *Sub-Cellular Biochemistry* 52 (2011), pp. 261–277.

[67] Michael G. Rosenfeld, Victoria V. Lunyak, and Christopher K. Glass. *Sensors and signals: A coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response*. 2006.

[68] Seth Frietze and Peggy J. Farnham. "Transcription Factor Effector Domains". In: *Sub-Cellular Biochemistry*. Vol. 52. Springer New York, 2011, pp. 261–277.

[69] Michael H. Kagey, Jamie J. Newman, Steve Bilodeau, Ye Zhan, David A. Orlando, Nynke L. Van Berkum, Christopher C. Ebmeier, Jesse Goossens, Peter B. Rahl, Stuart S. Levine, Dylan J. Taatjes, Job Dekker, and Richard A. Young. "Mediator and cohesin connect gene expression and chromatin architecture". In: *Nature* 467.7314 (2010), pp. 430–435.

[70] Axel Visel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M. Rubin, and Len A. Pennacchio. "ChIP-seq accurately predicts tissue-specific activity of enhancers". In: *Nature* 457.7231 (2009), pp. 854–858.

[71] Gerald Stampfel, Tomáš Kazmar, Olga Frank, Sebastian Wienerroither, Franziska Reiter, and Alexander Stark. "Transcriptional regulators form diverse groups with context-dependent regulatory functions." In: *Nature* 528.7580 (2015), pp. 147–51.

[72] Lisa Ann Cirillo, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S. Zaret. "Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4". In: *Molecular Cell* 9.2 (2002), pp. 279–289.

[73] Abdenour Soufi, Meilin Fernandez Garcia, Artur Jaroszewicz, Nebiyu Osman, Matteo Pellegrini, and Kenneth S. Zaret. "Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming". In: *Cell* 161.3 (2015), pp. 555–68.

[74] Michael P. Meers, Derek H. Janssens, and Steven Henikoff. "Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded". In: *Molecular Cell* 75.3 (2019), 562–575.e5.

[75] K. S. Zaret, J. M. Caravaca, A. Tulin, and T. Sekiya. "Nuclear mobility and mitotic chromosome binding similarities between pioneer transcription factor FoxA and linker histone H1". In: *Cold Spring Harbor Symposia on Quantitative Biology* 75 (2010), pp. 219–226.

[76] Hamish W. King and Robert J. Klose. "The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells". In: *eLife* 6 (2017).

[77] Roger D. Kornberg. "Chromatin structure: A repeating unit of histones and DNA". In: *Science* 184.4139 (1974), pp. 868–871.

[78] Guo Cheng Yuan, Yuen Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. "Molecular biology: Genome-scale identification of nucleosome positions in S. cerevisiae". In: *Science* 309.5734 (2005), pp. 626–630.

[79] Moyra Lawrence, Sylvain Daujat, and Robert Schneider. "Lateral Thinking: How Histone Modifications Regulate Gene Expression". In: *Trends in Genetics* 32.1 (2016), pp. 42–56.

[80] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. "Crystal structure of the nucleosome core particle at 2.8 A resolution". In: *Nature* 389.6648 (1997), pp. 251–260.

[81] Vikram Alva, Moritz Ammelburg, Johannes Söding, and Andrei N. Lupas. "On the origin of the histone fold". In: *BMC Structural Biology* 7.1 (2007), p. 17.

[82] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.4 (2019), pp. 207–220.

[83] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome". In: *Nature Genetics* 39.3 (2007), pp. 311–318.

[84] Helena Santos-Rosa, Robert Schneider, Andrew J. Bannister, Julia Sherriff, Bradley E. Bernstein, N. C.Tolga Emre, Stuart L. Schreiber, Jane Mellor, and Tony Kouzarides. "Active genes are tri-methylated at K4 of histone H3". In: *Nature* 419.6905 (2002), pp. 407–411.

[85] N Suka, Y Suka, A A Carmen, J Wu, and M Grunstein. "Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin." In: *Molecular cell* 8.2 (2001), pp. 473–9.

[86] Stephen Rea, Frank Eisenhaber, Dónal O'Carroll, Brian D. Strahl, Zu Wen Sun, Manfred Schmid, Susanne Opravil, Karl Mechtier, Chris P. Ponting, C. David Allis, and Thomas Jenuwein. "Regulation of chromatin structure by site-specific histone H3 methyltransferases". In: *Nature* 406.6796 (2000), pp. 593–599.

[87] Kangling Zhang, Hui Tang, Lan Huang, James W. Blankenship, Patrick R. Jones, Fan Xiang, Peter M. Yau, and Alma L. Burlingame. "Identification of acetylation and methylation sites of histone H3 from chicken erythrocytes by high-accuracy matrix-assisted laser desorption ionization-time-of-flight, matrix-assisted laser desorption ionization-postsource decay, and nanoelectrospray ion". In: *Analytical Biochemistry* 306.2 (2002), pp. 259–269.

[88] Dirk Schübeler. "Function and information content of DNA methylation". In: *Nature* 517.7534 (2015), pp. 321–326.

[89] Maxim V.C. Greenberg and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease". In: *Nature Reviews Molecular Cell Biology* 20.10 (2019), pp. 590–607.

[90] D. Schubeler. "Epigenetic Islands in a Genetic Ocean". In: *Science* 338.6108 (2012), pp. 756–757.

[91] Michael B Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J Oakeley, Dimos Gaidatzis, Vijay K Tiwari, and Dirk Schübeler. "DNA-binding factors shape the mouse methylome at distal regulatory regions." In: *Nature* 480.7378 (2011), pp. 490–5.

[92] Kasper Dindler Rasmussen and Kristian Helin. "Role of TET enzymes in DNA methylation, development, and cancer". In: *Genes & Development* 30.7 (2016), pp. 733–750.

[93] Peter A Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492.

[94] Zachary D. Smith and Alexander Meissner. "DNA methylation: roles in mammalian development". In: *Nature Reviews Genetics* 14.3 (2013), pp. 204–220.

[95] Heng Zhu, Guohua Wang, and Jiang Qian. "Transcription factors as readers and effectors of DNA methylation". In: *Nature Reviews Genetics* 17.9 (2016), pp. 551–565.

[96] Alexandre Hofer, Zheng J. Liu, and Shankar Balasubramanian. “Detection, Structure and Function of Modified DNA Bases”. In: *Journal of the American Chemical Society* 141.16 (2019), pp. 6420–6429.

[97] Ankur Jai Sood, Coby Viner, and Michael M. Hoffman. “DNAmod: the DNA modification database”. In: *Journal of Cheminformatics* 11.1 (2019), p. 30.

[98] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. “Multiscale 3D Genome Rewiring during Mouse Neural Development”. In: *Cell* 171.3 (2017), 557–572.e24.

[99] M. Jordan Rowley and Victor G. Corces. “Organizational principles of 3D genome architecture”. In: *Nature Reviews Genetics* 19.12 (2018), pp. 789–800.

[100] Eric D. Rubio, David J. Reiss, Piri L. Welcsh, Christine M. Disteche, Galina N. Filippova, Nitin S. Baliga, Ruedi Aebersold, Jeffrey A. Ranish, and Anton Krumm. “CTCF physically links cohesin to chromatin”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.24 (2008), pp. 8309–8314.

[101] Abraham S Weintraub, Charles H Li, Alicia V Zamudio, Alla A Sigova, Nancy M Hannett, Daniel S Day, Brian J Abraham, Malkiel A Cohen, Behnam Nabet, Dennis L Buckley, Yang Eric Guo, Denes Hnisz, Rudolf Jaenisch, James E Bradner, Nathanael S Gray, and Richard A Young. “YY1 Is a Structural Regulator of Enhancer-Promoter Loops.” In: *Cell* 171.7 (2017), 1573–1588.e28.

[102] Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”. In: *Science* 326.5950 (2009), pp. 289–293.

[103] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. “Topological domains in mammalian genomes identified by analysis of chromatin interactions.” In: *Nature* 485.7398 (2012), pp. 376–80.

[104] Jill M. Dowen, Zi Peng Fan, Denes Hnisz, Gang Ren, Brian J. Abraham, Lyndon N. Zhang, Abraham S. Weintraub, Jurian Schuijers, Tong Ihn Lee, Keji Zhao, and Richard A. Young. “Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes”. In: *Cell* 159.2 (2014), pp. 374–387.

[105] Eileen E.M. Furlong and Michael Levine. “Developmental enhancers and chromosome topology”. In: *Science* 361.6409 (2018), pp. 1341–1345.

[106] Denes Hnisz, Daniel S Day, and Richard A Young. “Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control.” In: *Cell* 167.5 (2016), pp. 1188–1200.

[107] Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. “Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor”. In: *Cell* 149.6 (2012), pp. 1233–1244.

[108] Stefanie L. Morgan, Natasha C. Mariano, Abel Bermudez, Nicole L. Arruda, Fangting Wu, Yunhai Luo, Gautam Shankar, Lin Jia, Huiling Chen, Ji Fan Hu, Andrew R. Hoffman, Chiao Chain Huang, Sharon J. Pitteri, and Kevin C. Wang. "Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping". In: *Nature Communications* 8.1 (2017), pp. 1–9.

[109] Nezha S. Benabdallah, Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. "Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation". In: *Molecular Cell* 76.3 (2019), 473–484.e7.

[110] Terrence S Furey. "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions". In: *Nature Reviews Genetics* 13.12 (2012), pp. 840–852.

[111] Peter J Park. "ChIP–seq: advantages and challenges of a maturing technology". In: *Nature Reviews Genetics* 10.10 (2009), pp. 669–680.

[112] Laura Baranello, Fedor Kouzine, Suzanne Sanford, and David Levens. "ChIP bias as a function of cross-linking time." In: *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 24.2 (2016), pp. 175–81.

[113] Sivakanthan Kasinathan, Guillermo A. Orsi, Gabriel E. Zentner, Kami Ahmad, and Steven Henikoff. "High-resolution mapping of transcription factor binding sites on native chromatin". In: *Nature Methods* 11.2 (2014), pp. 203–209.

[114] Nicola Festuccia, Nick Owens, Thaleia Papadopoulou, Inma Gonzalez, Alexandra Tachtsidi, Sandrine Vandoermel-Pournin, Elena Gallego, Nancy Gutierrez, Agnès Dubois, Michel Cohen-Tannoudji, and Pablo Navarro. "Transcription factor activity and nucleosome organization in mitosis". In: *Genome Research* 29.2 (2019), pp. 250–260.

[115] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit. "Chromatin state dynamics during blood formation". In: *Science* 345.6199 (2014), pp. 943–9.

[116] Christian Schmidl, André F Rendeiro, Nathan C Sheffield, and Christoph Bock. "ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors". In: *Nature Methods* 12.10 (2015), pp. 963–5.

[117] Ho Sung Rhee and B. Franklin Pugh. "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution". In: *Cell* 147.6 (2011), pp. 1408–1419.

[118] Qiye He, Jeff Johnston, and Julia Zeitlinger. "ChIP-nexus enables improved detection of in vivo transcription factor binding footprints". In: *Nature Biotechnology* 33.4 (2015), pp. 395–401.

[119] Peter J Skene and Steven Henikoff. "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites". In: *eLife* 6 (2017), e21856.

[120] Sarah J. Hainer, Ana Bošković, Kurtis N. McCannell, Oliver J. Rando, and Thomas G. Fazzio. "Profiling of Pluripotency Factors in Single Cells and Early Embryos". In: *Cell* 177.5 (2019), 1319–1329.e11.

[121] Stefan Canzar and Steven L. Salzberg. “Short Read Mapping: An Algorithmic Tour”. In: *Proceedings of the IEEE* 105.3 (2017), pp. 436–458.

[122] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359.

[123] Aaron Diaz, Kiyoub Park, Daniel A Lim, and Jun S Song. “Normalization, bias correction, and peak calling for ChIP-seq”. In: *Statistical Applications in Genetics and Molecular Biology* 11.3 (2012), Article 9.

[124] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A Grüning, and Thomas Manke. “deepTools: a flexible platform for exploring deep-sequencing data”. In: *Nucleic Acids Research* 42.W1 (2014), W187–W191.

[125] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. “Identifying ChIP-seq enrichment using MACS”. In: *Nature Protocols* 7.9 (2012), pp. 1728–1740.

[126] Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard. “Features that define the best ChIP-seq peak calling algorithms”. In: *Briefings in Bioinformatics* 18.3 (2016), bbw035.

[127] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. “Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities”. In: *Molecular Cell* 38.4 (2010), pp. 576–589.

[128] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. “GREAT improves functional interpretation of cis-regulatory regions”. In: *Nature Biotechnology* 28.5 (2010), pp. 495–501.

[129] Alan P. Boyle, Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. “High-Resolution Mapping and Characterization of Open Chromatin across the Genome”. In: *Cell* 132.2 (2008), pp. 311–322.

[130] Paul G. Giresi, Jonghwan Kim, Ryan M. McDaniell, Vishwanath R. Iyer, and Jason D. Lieb. “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin”. In: *Genome Research* 17.6 (2007), pp. 877–885.

[131] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature Methods* 10.12 (2013), pp. 1213–1218.

[132] Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis”. In: *Genome Biology* 21.1 (2020), p. 22.

[133] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. “Single-cell chromatin accessibility reveals principles of regulatory variation.” In: *Nature* (2015).

[134] Zhijian Li, Marcel H. Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G. Costa. “Identification of transcription factor binding sites using ATAC-seq”. In: *Genome Biology* 20.1 (2019), p. 45.

[135] Asllhan Karabacak Calviello, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. “Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling”. In: *Genome Biology* 20.1 (2019), p. 42.

[136] Alicia N Schep, Jason D Buenrostro, Sarah K Denny, Katja Schwartz, Gavin Sherlock, and William J Greenleaf. “Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions”. In: *Genome Research* 25.11 (2015), pp. 1757–1770.

[137] Thomas A. Down, Vardhman K. Rakyan, Daniel J. Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Gräf, Nathan Johnson, Javier Herrero, Eleni M. Tomazou, Natalie P. Thorne, Liselotte Bäckdahl, Marlis Herberth, Kevin L. Howe, David K. Jackson, Marcos M. Miretti, John C. Marioni, Ewan Birney, Tim J.P. Hubbard, Richard Durbin, Simon Tavaré, and Stephan Beck. “A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis”. In: *Nature Biotechnology* 26.7 (2008), pp. 779–785.

[138] Theresa K Kelly, Yaping Liu, Fides D Lay, Gangning Liang, Benjamin P Berman, and Peter A Jones. “Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules”. In: *Genome Research* 22.12 (2012), pp. 2497–2506.

[139] Felix Krueger and Simon R. Andrews. “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications”. In: *Bioinformatics* 27.11 (2011), pp. 1571–1572.

[140] Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. “Identification of active regulatory regions from DNA methylation data”. In: *Nucleic Acids Research* 41.16 (2013), e155–e155.

[141] Stefan Schoenfelder and Peter Fraser. “Long-range enhancer–promoter contacts in gene expression control”. In: *Nature Reviews Genetics* 20.8 (2019), pp. 437–455.

[142] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. “Capturing chromosome conformation”. In: *Science* 295.5558 (2002), pp. 1306–1311.

[143] Anthony D. Schmitt, Ming Hu, and Bing Ren. “Genome-wide mapping and analysis of chromosome architecture”. In: *Nature Reviews Molecular Cell Biology* 17.12 (2016), pp. 743–755.

[144] Stefan Schoenfelder, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola-Maria Javierre, Takashi Nagano, Yulia Katsman, Moorthy Sakthidevi, Steven W Wingett, Emilia Dimitrova, Andrew Dimond, Lucas B Edelman, Sarah Elderkin, Kristina Tabbada, Elodie Darbo, Simon Andrews, Bram Herman, Andy Higgs, Emily LeProust, Cameron S Osborne, Jennifer A Mitchell, Nicholas M Luscombe, and Peter Fraser. “The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements”. In: *Genome Research* 25.4 (2015), pp. 582–597.

[145] Onkar Joshi, Shuang-Yin Wang, Tatyana Kuznetsova, Yaser Atlasi, Tianran Peng, Pierre J. Fabre, Ehsan Habibi, Jani Shaik, Sadia Saeed, Lusy Handoko, Todd Richmond, Mikhail Spivakov, Daniel Burgess, and Hendrik G. Stunnenberg. "Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency". In: *Cell Stem Cell* 17.6 (2015), pp. 748–757.

[146] Pelin Sahlén, Ilgar Abdullayev, Daniel Ramsköld, Liudmila Matskova, Nemanja Rilakovic, Britta Lötstedt, Thomas J. Albert, Joakim Lundeberg, and Rickard Sandberg. "Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution". In: *Genome Biology* 16.1 (2015), p. 156.

[147] Steven Wingett, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. "HiCUP: Pipeline for mapping and processing Hi-C data". In: *F1000Research* 4 (2015), p. 1310.

[148] Jonathan Cairns, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser, and Mikhail Spivakov. "CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data". In: *Genome Biology* 17.1 (2016), p. 127.

[149] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11 (2019), pp. 631–656.

[150] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. "The transcriptional landscape of the yeast genome defined by RNA sequencing". In: *Science* 320.5881 (2008), pp. 1344–1349.

[151] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7 (2008), pp. 621–628.

[152] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[153] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (2016), pp. 525–527.

[154] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4 (2017), pp. 417–419.

[155] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. "A survey of best practices for RNA-seq data analysis". In: *Genome Biology* 17.1 (2016), p. 13.

[156] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014), p. 550.

[157] L. Stirling Churchman and Jonathan S. Weissman. "Nascent transcript sequencing visualizes transcription at nucleotide resolution". In: *Nature* 469.7330 (2011), pp. 368–373.

[158] Bjoern Schwalb, Margaux Michel, Benedikt Zacher, Katja Frü Hauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. "TT-seq maps the human transient transcriptome". In: *Science* 352.6290 (2016), pp. 1225–1228.

[159] Charles G. Danko, Nasun Hah, Xin Luo, André L. Martins, Leighton Core, John T. Lis, Adam Siepel, and W. Lee Kraus. "Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells". In: *Molecular Cell* 50.2 (2013), pp. 212–222.

[160] Charles G Danko, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. "Identification of active transcriptional regulatory elements from GRO-seq data". In: *Nature Methods* 12.5 (2015), pp. 433–8.

[161] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. "Transcriptional enhancers: from properties to genome-wide predictions". In: *Nature Reviews Genetics* 15.4 (2014), pp. 272–286.

[162] Rui R. Catarino and Alexander Stark. "Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation". In: *Genes & Development* 32.3-4 (2018), pp. 202–223.

[163] Molly Gasperini, Jacob M. Tome, and Jay Shendure. "Towards a comprehensive catalogue of validated and target-linked human enhancers". In: *Nature Reviews Genetics* (2020).

[164] Benjamin P. Berman, Yutaka Nibu, Barret D. Pfeiffer, Pavel Tomancak, Susan E. Celniker, Michael Levine, Gerald M. Rubin, and Michael B. Eisen. "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.2 (2002), pp. 757–762.

[165] Outi Hallikas, Kimmo Palin, Natalia Sinjushina, Reetta Rautiainen, Juha Partanen, Esko Ukkonen, and Jussi Taipale. "Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity". In: *Cell* 124.1 (2006), pp. 47–59.

[166] Shyam Prabhakar, Francis Poulin, Malak Shoukry, Veena Afzal, Edward M. Rubin, Olivier Couronne, and Len A. Pennacchio. "Close sequence comparisons are sufficient to identify human cis-regulatory elements". In: *Genome Research* 16.7 (2006), pp. 855–863.

[167] N. C. Sheffield, R. E. Thurman, L. Song, A. Safi, J. A. Stamatoyannopoulos, B. Lenhard, G. E. Crawford, and T. S. Furey. "Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions". In: *Genome Research* 23.5 (2013), pp. 777–788.

[168] Jeff Vierstra, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R Scott Hansen, Sandra Stehling-Sun, Peter J Sabo, Rachel Byron, Richard Humbert, Robert E Thurman, Audra K Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Erika Giste, Eric Haugen, Douglas Dunn, Matthew S Wilken, Steven Josefowicz, Robert Samstein, K.-H. Chang, Evan E Eichler, Marella De Bruijn, Thomas A Reh, Arthur Skoultchi, Alexander Rudensky, Stuart H Orkin, Thalia Papayannopoulou, Piper M Treuting, Licia Selleri, Rajinder Kaul, Mark Groudine, M A Bender, and John A Stamatoyannopoulos. "Mouse regulatory DNA

landscapes reveal global principles of cis-regulatory evolution". In: *Science* 346.6212 (2014), pp. 1007–1012.

[169] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, Ravindra Majeti, and Howard Y Chang. "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution". In: *Nature Genetics* (2016).

[170] Robin Andersson, Albin Sandelin, and Charles G Danko. "A unified architecture of transcriptional regulatory elements". In: *Trends in Genetics* 31.8 (2015), pp. 426–433.

[171] Robin Andersson. "Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model." In: *BioEssays: news and reviews in molecular, cellular and developmental biology* 37.3 (2015), pp. 314–23.

[172] Robert S. Young, Yatendra Kumar, Wendy A. Bickmore, and Martin S. Taylor. "Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers". In: *Genome Biology* 18.1 (2017), p. 242.

[173] Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, L. M. Boryn, Martina Rath, and Alexander Stark. "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq". In: *Science* 339.6123 (2013), pp. 1074–1077.

[174] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G. Callan, Justin B. Kinney, Manolis Kellis, Eric S. Lander, and Tarjei S. Mikkelsen. "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay". In: *Nature Biotechnology* 30.3 (2012), pp. 271–277.

[175] Jamie C Kwasnieski, Ilaria Mogno, Connie A Myers, Joseph C Corbo, and Barak A Cohen. "Complex effects of nucleotide variants in a mammalian cis-regulatory element". In: *Proceedings of the National Academy of Sciences* 109.47 (2012), pp. 19498–19503.

[176] Felix Muerdter, Łukasz M Boryń, Ashley R Woodfin, Christoph Neumayr, Martina Rath, Muhammad A Zabidi, Michaela Pagani, Vanja Haberle, Tomáš Kazmar, Rui R Catarino, Katharina Schernhuber, Cosmas D Arnold, and Alexander Stark. "Resolving systematic errors in widely used enhancer activity assays in human cells". In: *Nature Methods* (2017).

[177] Laurent Vanhille, Aurélien Griffon, Muhammad Ahmad Maqbool, Joaquin Zacarias-Cabeza, Lan T.M. Dao, Nicolas Fernandez, Benoit Ballester, Jean Christophe Andrau, and Salvatore Spicuglia. "High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq". In: *Nature Communications* 6.1 (2015), p. 6905.

[178] Christopher M Vockley, Anthony M D'Ippolito, Ian C McDowell, William H Majoros, Alexias Safi, Lingyun Song, Gregory E Crawford, and Timothy E Reddy. "Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome." In: *Cell* 166.5 (2016), 1269–1281.e19.

[179] Xinchen Wang, Liang He, Sarah M. Goggin, Alham Saadat, Li Wang, Nasa Sinnott-Armstrong, Melina Claussnitzer, and Manolis Kellis. "High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human". In: *Nature Communications* 9.1 (2018), p. 5380.

[180] Tahsin Stefan Barakat, Florian Halbritter, Man Zhang, André F Rendeiro, Elena Perenthaler, Christoph Bock, and Ian Chambers. "Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells". In: *Cell Stem Cell* 23.2 (2018), 276–288.e8.

[181] Fumitaka Inoue, Martin Kircher, Beth Martin, Gregory M Cooper, Daniela M Witten, Michael T McManus, Nadav Ahituv, and Jay Shendure. "A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity." In: *Genome research* 27.1 (2017), pp. 38–52.

[182] Tong Ihn Lee and Richard A Young. "Transcriptional regulation and its misregulation in disease." In: *Cell* 152.6 (2013), pp. 1237–51.

[183] Mo Li and Juan Carlos Izpisua Belmonte. "Ground rules of the pluripotency gene regulatory network". In: *Nature Reviews Genetics* 18.3 (2017), pp. 180–191.

[184] Christoph Neumayr, Michaela Pagani, Alexander Stark, and Cosmas D. Arnold. "STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries". In: *Current Protocols in Molecular Biology* 128.1 (2019), e105.

[185] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137.

[186] A. Akalin, V. Franke, K. Vlahovi ek, C. E. Mason, and D. Schubeler. "genomation: a toolkit to summarize, annotate and visualize genomic intervals". In: *Bioinformatics* 31.7 (2015), pp. 1127–1129.

[187] Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (2014), pp. 923–930.

[188] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. "Measuring reproducibility of high-throughput experiments". In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1752–1779. arXiv: *1110.4705*.

[189] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[190] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports* 9.1 (2019), p. 9354.

[191] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. "Software for Computing and Annotating Genomic Ranges". In: *PLOS Computational Biology* 9.8 (2013).

[192] Nathan C Sheffield and Christoph Bock. "LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor". In: *Bioinformatics* 32.4 (2016), pp. 587–589.

[193] Rongbin Zheng, Changxin Wan, Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, Chen Hao Chen, Myles Brown, Xiaoyan Zhang, Clifford A. Meyer, and X. Shirley Liu. "Cistrome Data Browser: Expanded datasets and new tools for gene regulatory analysis". In: *Nucleic Acids Research* 47.D1 (2019), pp. D729–D735.

[194] Alaa Tharwat. "Classification assessment methods". In: *Applied Computing and Informatics* (2018).

[195] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[196] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization". In: *Bioinformatics* 31.14 (2015), pp. 2382–2383.

[197] Jonathan M Mudge and Jennifer Harrow. "Creating reference gene annotation for the mouse C57BL6/J genome assembly." In: *Mammalian genome : official journal of the International Mammalian Genome Society* 26.9-10 (2015), pp. 366–78.

[198] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. "High-resolution TADs reveal DNA sequences underlying genome organization in flies". In: *Nature Communications* 9.1 (2018), p. 189.

[199] Yaser Atlasi, Wout Megchelenbrink, Tianran Peng, Ehsan Habibi, Onkar Joshi, Shuang-Yin Wang, Cheng Wang, Colin Logie, Ina Poser, Hendrik Marks, and Hendrik G. Stunnenberg. "Epigenetic modulation of a hardwired 3D chromatin landscape in two naive states of pluripotency". In: *Nature Cell Biology* (2019), p. 1.

[200] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters". In: *OMICS: A Journal of Integrative Biology* 16.5 (2012), pp. 284–287.

[201] Denise N Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, Jacob Windsor, Nuno Nunes, Jonathan Mélius, Elisa Cirillo, Susan L Coort, Daniela Digles, Friederike Ehrhart, Pieter Giesbertz, Marianthi Kalafati, Marvin Martens, Ryan Miller, Kozo Nishida, Linda Rieswijk, Andra Waagmeester, Lars M T Eijssen, Chris T Evelo, Alexander R Pico, and Egon L Willighagen. "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research". In: *Nucleic Acids Research* 46.D1 (2018), pp. D661–D667.

[202] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. "The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms". In: *Nucleic Acids Research* 45.D1 (2017), pp. D51–D55.

[203] D. Karolchik. "The UCSC Table Browser data retrieval tool". In: *Nucleic Acids Research* 32.90001 (2004), pp. 493D–496.

[204] Mahmoud-Reza Rafiee, Charles Girardot, Gianluca Sigismondo, and Jeroen Krijgsveld. "Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins". In: *Molecular Cell* 64.3 (2016), pp. 624–635.

[205] Guido van Mierlo, René A.M. Dirks, Laura De Clerck, Arie B. Brinkman, Michelle Huth, Susan L. Kloet, Nehmé Saksouk, Leonie I. Kroeze, Sander Willems, Matthias Farlik, Christoph Bock, Joop H. Jansen, Dieter Deforce, Michiel Vermeulen, Jérôme Déjardin, Maarten Dhaenens, and Hendrik Marks. "Integrative Proteomic Profiling Reveals PRC2-Dependent Epigenetic Crosstalk Maintains Ground-State Pluripotency". In: *Cell Stem Cell* 24.1 (2019), 123–137.e8.

[206] Max Kuhn. "Building predictive models in R using the caret package". In: *Journal of Statistical Software* 28.5 (2008), pp. 1–26.

[207] Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and G B M Developers. *gbm: Generalized Boosted Regression Models*. 2019.

[208] Mahmoud Ghandi, Morteza Mohammad-Noori, Narges Ghareghani, Dongwon Lee, Levi Garraway, and Michael A. Beer. "gkmSVM: an R package for gapped-kmer SVM". In: *Bioinformatics* 32.14 (2016), pp. 2205–2207.

[209] J Bergstra, D Yamins, and D D Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". In: *ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning* 28 (2013), pp. 115–123.

[210] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* (2019), p. 1.

[211] Alexey Natekin and Alois Knoll. "Gradient boosting machines, a tutorial". In: *Frontiers in Neurorobotics* 7 (2013).

[212] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: *arXiv* (2016). arXiv: *1605.01713*.

[213] Avanti Shrikumar, Katherine Tian, Anna Shcherbina, Žiga Avsec, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. "TF-MoDISco v0.4.2.2-alpha: Technical Note". In: *arXiv* (2018). arXiv: *1811.00416*.

[214] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. "Quantifying similarity between motifs". In: *Genome Biology* 8.2 (2007).

[215] Shaun Mahony and Panayiotis V Benos. "STAMP: a web tool for exploring DNA-binding motif similarities". In: *Nucleic Acids Research* 35.Web Server (2007), W253–W258.

[216] Christina Galonska, Michael J. Ziller, Rahul Karnik, and Alexander Meissner. "Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming". In: *Cell Stem Cell* 17.4 (2015), pp. 462–470.

[217] Clara Lopes Novo, Biola-Maria Javierre, Jonathan Cairns, Anne Segonds-Pichon, Steven W Wingett, Paula Freire-Pritchett, Mayra Furlan-Magaril, Stefan Schoenfelder, Peter Fraser, and Peter J Rugg-Gunn. "Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition". In: *Cell Reports* 22.10 (2018), pp. 2615–2627.

[218] Clifford A. Meyer and X. Shirley Liu. “Identifying and mitigating bias in next-generation sequencing methods for chromatin biology”. In: *Nature Reviews Genetics* 15.11 (2014), pp. 709–721.

[219] Cynthia A. Kalita, Christopher D. Brown, Andrew Freiman, Jenna Isherwood, Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. “High-throughput characterization of genetic effects on DNA-protein binding and gene transcription”. In: *Genome Research* 28.11 (2018), pp. 1701–1708.

[220] Nicola Festuccia, Rodrigo Osorno, Valerie Wilson, and Ian Chambers. “The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states”. In: *Current Opinion in Genetics & Development* 23.5 (2013), pp. 504–511.

[221] Patricia Respuela, Miloš Nikolić, Minjia Tan, Peter Frommolt, Yingming Zhao, Joanna Wysocka, and Alvaro Rada-Iglesias. “Foxd3 Promotes Exit from Naive Pluripotency through Enhancer Decommissioning and Inhibits Germline Specification”. In: *Cell Stem Cell* 18.1 (2016), pp. 118–133.

[222] Bony De Kumar, Hugo J Parker, Mark E Parrish, Jeffrey J Lange, Brian D Slaughter, Jay R Unruh, Ariel Paulson, and Robb Krumlauf. “Dynamic regulation of Nanog and stem cell-signaling pathways by Hoxa1 during early neuro-ectodermal differentiation of ES cells”. In: *Proceedings of the National Academy of Sciences* 114.23 (2017), pp. 5838–5845.

[223] Steven Blinka and Sridhar Rao. “Nanog Expression in Embryonic Stem Cells - An Ideal Model System to Dissect Enhancer Function”. In: *BioEssays* (2017), p. 1700086.

[224] Nicola A Kearns, Hannah Pham, Barbara Tabak, Ryan M Genga, Noah J Silverstein, Manuel Garber, and René Maehr. “Functional annotation of native enhancers with a Cas9–histone demethylase fusion”. In: *Nature Methods* 12.5 (2015), pp. 401–403.

[225] Leehee Weinberger, Muneef Ayyash, Noa Novershtern, and Jacob H. Hanna. “Dynamic stem cell states: naive to primed pluripotency in rodents and humans”. In: *Nature Reviews Molecular Cell Biology* 17.3 (2016), pp. 155–169.

[226] Dario Acampora, Luca G Di Giovannantonio, and Antonio Simeone. “Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition”. In: *Development* 140.1 (2013), pp. 43–55.

[227] Amandine Chatagnon, Philippe Veber, Valérie Morin, Justin Bedo, Gérard Triqueneaux, Marie Sémon, Vincent Laudet, Florence D’Alché-Buc, and Gérard Benoit. “RAR/RXR binding dynamics distinguish pluripotency from differentiation associated cis-regulatory elements”. In: *Nucleic Acids Research* 43.10 (2015), pp. 4833–4854.

[228] Zahra Anvar, Marco Cammisa, Vincenzo Riso, Ilaria Baglivo, Harpreet Kukreja, Angela Sparago, Michael Girardot, Shraddha Lad, Italia De Feis, Flavia Cerrato, Claudia Angelini, Robert Feil, Paolo V. Pedone, Giovanna Grimaldi, and Andrea Riccio. “ZFP57 recognizes multiple and closely spaced sequence motif variants to maintain repressive epigenetic marks in mouse embryonic stem cells”. In: *Nucleic Acids Research* 44.3 (2016), pp. 1118–1132.

[229] Hong Sook Kim, Yuliang Tan, Wubin Ma, Daria Merkurjev, Eugin Destici, Qi Ma, Tom Suter, Kenneth Ohgi, Meyer Friedman, Dorota Skowronska-Krawczyk, and Michael G. Rosenfeld. "Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells". In: *Nature* (2018), p. 1.

[230] Eliezer Calo and Joanna Wysocka. "Modification of Enhancer Chromatin: What, How, and Why?" In: *Molecular Cell* 49.5 (2013), pp. 825–837.

[231] Sara Cruz-Molina, Patricia Respuela, Christina Tebartz, Petros Kolovos, Milos Nikolic, Raquel Fueyo, Wilfred F.J. van Ijcken, Frank Grosveld, Peter Frommolt, Hisham Bazzi, and Alvaro Rada-Iglesias. "PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation". In: *Cell Stem Cell* 20.5 (2017), 689–705.e9.

[232] Sara Martire, Aishwarya A. Gogate, Amanda Whitmill, Amanuel Tafessu, Jennifer Nguyen, Yu-Ching Teng, Melodi Tastemel, and Laura A. Banaszynski. "Phosphorylation of histone H3.3 at serine 31 promotes p300 activity and enhancer acetylation". In: *Nature Genetics* 51.6 (2019), pp. 941–946.

[233] Michael Girardot, Ryutaro Hirasawa, Salim Kacem, Lauriane Fritsch, Julien Pontis, Satya K. Kota, Doria Filipponi, Eric Fabbrizio, Claude Sardet, Felix Lohmann, Shilpa Kadam, Slimane Ait-Si-Ali, and Robert Feil. "PRMT5-mediated histone H4 arginine-3 symmetrical dimethylation marks chromatin at G + C-rich regions of the mouse genome". In: *Nucleic Acids Research* 42.1 (2014), pp. 235–248.

[234] Taiping Chen and Sharon Y R Dent. "Chromatin modifiers and remodellers: regulators of cellular differentiation". In: *Nature Reviews Genetics* 15.2 (2014), pp. 93–106.

[235] Ryan Raisner, Samir Kharbanda, Lingyan Jin, Edwin Jeng, Emily Chan, Mark Merchant, Peter M. Haverty, Russell Bainer, Tommy Cheung, David Arnott, E. Megan Flynn, F. Anthony Romero, Steven Magnuson, and Karen E. Gascoigne. "Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation". In: *Cell Reports* 24.7 (2018), pp. 1722–1729.

[236] Nergiz Dogan, Weisheng Wu, Christapher S. Morrissey, Kuan-Bei Chen, Aaron Stonestrom, Maria Long, Cheryl A. Keller, Yong Cheng, Deepti Jain, Axel Visel, Len A. Pennacchio, Mitchell J. Weiss, Gerd A. Blobel, and Ross C. Hardison. "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility". In: *Epigenetics & Chromatin* 8.1 (2015), p. 16.

[237] Ping-Hsing Tsai, Yueh Chien, Mong-Lien Wang, Chih-Hung Hsu, Benoit Laurent, Shih-Jie Chou, Wei-Chao Chang, Chian-Shiu Chien, Hsin-Yang Li, Hsin-Chen Lee, Teh-Ia Huo, Jui-Hung Hung, Chung-Hsuan Chen, and Shih-Hwa Chiou. "Ash2l interacts with Oct4-stemness circuitry to promote super-enhancer-driven pluripotency network". In: *Nucleic Acids Research* 47.19 (2019), pp. 10115–10133.

[238] Raffaella Di Micco, Barbara Fontanals-Cirera, Vivien Low, Panagiotis Ntziachristos, Stephanie K. Yuen, Claudia D. Lovell, Igor Dolgalev, Yoshiya Yonekubo, Guangtao Zhang, Elena Rusinova, Guillermo Gerona-Navarro, Marta Cañamero, Michael Ohlmeyer, Iannis Aifantis, Ming Ming Zhou, Aristotelis Tsirigos, and Eva Hernando. "Control of embryonic stem cell identity by brd4-dependent transcriptional elonga-

tion of super-enhancer-associated pluripotency genes". In: *Cell Reports* 9.1 (2014), pp. 234–247.

[239] Richard A. Young. "Control of the Embryonic Stem Cell State". In: *Cell* 144.6 (2011), pp. 940–954.

[240] Shereen Jamaladdin, Richard D W Kelly, Laura O'Regan, Oliver M Dovey, Grace E Hodson, Christopher J Millard, Nicola Portolano, Andrew M Fry, John W R Schwabe, and Shaun M Cowley. "Histone deacetylase (HDAC) 1 and 2 are essential for accurate cell division and the pluripotency of embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 111.27 (2014), pp. 9840–9845.

[241] Erin A Clark, Feizhen Wu, Yirui Chen, Paco Kang, Ursula B Kaiser, Rui Fang, and Yujiang G Shi. "GR and LSD1/KDM1A-Targeted Gene Activation Requires Selective H3K4me2 Demethylation at Enhancers". In: *Cell Reports* 27.12 (2019), 3522–3532.e3.

[242] Diego Pasini, Paul A C Cloos, Julian Walfridsson, Linda Olsson, John-Paul Bukowski, Jens V Johansen, Mads Bak, Niels Tommerup, Juri Rappsilber, and Kristian Helin. "JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells." In: *Nature* 464.7286 (2010), pp. 306–10.

[243] Kavitha Sarma, Catherine Cifuentes-Rojas, Ayla Ergun, Amanda Del Rosario, Yesu Jeon, Forest White, Ruslan Sadreyev, and Jeannie T Lee. "ATRX Directs Binding of PRC2 to Xist RNA and Polycomb Targets." In: *Cell* 159.5 (2014), p. 1228.

[244] Jin He, Li Shen, Ma Wan, Olena Taranova, Hao Wu, and Yi Zhang. "Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes". In: *Nature Cell Biology* 15.4 (2013), pp. 373–384.

[245] Anca M Farcas, Neil P Blackledge, Ian Sudbery, Hannah K Long, Joanna F McGouran, Nathan R Rose, Sheena Lee, David Sims, Andrea Cerase, Thomas W Sheahan, Haruhiko Koseki, Neil Brockdorff, Chris P Ponting, Benedikt M Kessler, and Robert J Klose. "KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands." In: *eLife* 1 (2012), e00205.

[246] Junling Jia, Xiaobin Zheng, Gangqing Hu, Kairong Cui, Junqi Zhang, Anying Zhang, Hao Jiang, Bingwen Lu, John Yates, Chengyu Liu, Keji Zhao, and Yixian Zheng. "Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning." In: *Cell* 151.3 (2012), pp. 576–89.

[247] Melodi Tastemel, Aishwarya A. Gogate, Venkat S. Malladi, Kim Nguyen, Courtney Mitchell, Laura A. Banaszynski, and Xiaoying Bai. "Transcription pausing regulates mouse embryonic stem cell differentiation". In: *Stem Cell Research* 25 (2017), pp. 250–255.

[248] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nature Genetics* 47.6 (2015), pp. 598–606.

[249] Zheng Wang, Efrat Oron, Brynna Nelson, Spiro Razis, and Natalia Ivanova. "Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells". In: *Cell Stem Cell* 10.4 (2012), pp. 440–454.

[250] Antti Lignell, Laura Kerosuo, Sebastian J Streichan, Long Cai, and Marianne E Bronner. “Identification of a neural crest stem cell niche by Spatial Genomic Analysis”. In: *Nature Communications* 8.1 (2017), p. 1830.

[251] Masamitsu Sone, Nobuhiro Morone, Tomonori Nakamura, Akito Tanaka, Keisuke Okita, Knut Woltjen, Masato Nakagawa, John E. Heuser, Yasuhiro Yamada, Shinya Yamanaka, and Takuya Yamamoto. “Hybrid Cellular Metabolism Coordinated by Zic3 and Esrrb Synergistically Enhances Induction of Naive Pluripotency”. In: *Cell Metabolism* 25.5 (2017), 1103–1117.e6.

[252] Jia-Chi Yeo, Jianming Jiang, Zi-Ying Tan, Guo-Rong Yim, Jia-Hui Ng, Jonathan Göke, Petra Kraus, Hongqing Liang, Kevin Andrew Uy Gonzales, Han-Chung Chong, Cheng-Peow Tan, Yee-Siang Lim, Nguan-Soon Tan, Thomas Lufkin, and Huck-Hui Ng. “Klf2 Is an Essential Factor that Sustains Ground State Pluripotency”. In: *Cell Stem Cell* 14.6 (2014), pp. 864–872.

[253] Fei Yi, Laura Pereira, Jackson A. Hoffman, Brian R. Shy, Courtney M. Yuen, David R. Liu, and Bradley J. Merrill. “Opposing effects of Tcf3 and Tcf1 control Wnt stimulation of embryonic stem cell self-renewal”. In: *Nature Cell Biology* 13.7 (2011), pp. 762–770.

[254] Liangqi Xie, Sharon E Torigoe, Jifang Xiao, Daniel H Mai, Li Li, Fred P Davis, Peng Dong, Herve Marie-Nelly, Jonathan Grimm, Luke Lavis, Xavier Darzacq, Claudia Cattoglio, Zhe Liu, and Robert Tjian. “A dynamic interplay of enhancer elements regulates Klf4 expression in naïve pluripotency”. In: *Genes & Development* 31.17 (2017), pp. 1795–1808.

[255] Nicola Festuccia, Agnès Dubois, Sandrine Vandormael-Pournin, Elena Gallego Tejeda, Adrien Mouren, Sylvain Bessonnard, Florian Mueller, Caroline Proux, Michel Cohen-Tannoudji, and Pablo Navarro. “Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network”. In: *Nature Cell Biology* 18.11 (2016), pp. 1139–1148.

[256] Jason Ernst, Alexandre Melnikov, Xiaolan Zhang, Li Wang, Peter Rogov, Tarjei S Mikkelsen, and Manolis Kellis. “Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions”. In: *Nature Biotechnology* 34.11 (2016), pp. 1180–1190.

[257] Harry Y Zhou, Yulia Katsman, Navroop K Dhaliwal, Scott Davidson, Neil N Macpherson, Moorthy Sakthidevi, Felicia Collura, and Jennifer A Mitchell. “A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential”. In: *Genes & Development* 28.24 (2014), pp. 2699–2711.

[258] Hitoshi Niwa. “The principles that govern transcription factor network functions in stem cells”. In: *Development* 145.6 (2018), dev157420.

[259] Victor Heurtier, Nick Owens, Inma Gonzalez, Florian Mueller, Caroline Proux, Damien Mornico, Philippe Clerc, Agnes Dubois, and Pablo Navarro. “The molecular logic of Nanog-induced self-renewal in mouse embryonic stem cells”. In: *Nature Communications* 10.1 (2019), p. 1109.

[260] Nicola Festuccia, Florian Halbritter, Andrea Corsinotti, Alessia Gagliardi, Douglas Colby, Simon R Tomlinson, and Ian Chambers. “Esrrb extinction triggers dismantling of naïve pluripotency and marks commitment to differentiation”. In: *The EMBO Journal* 37.21 (2018), e95476.

[261] Kenjiro Adachi, Wolfgang Kopp, Guangming Wu, Sandra Heising, Boris Greber, Martin Stehling, Marcos J. Araúzo-Bravo, Stefan T. Boerno, Bernd Timmermann, Martin Vingron, and Hans R. Schöler. "Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency". In: *Cell Stem Cell* 23.2 (2018), 266–275.e6.

[262] Akira Nishiyama, Alexei A Sharov, Yulan Piao, Misa Amano, Tomokazu Amano, Hien G Hoang, Bernard Y Binder, Richard Tapnio, Uwem Bassey, Justin N Malinou, Lina S Correa-Cerro, Hong Yu, Li Xin, Emily Meyers, Michal Zalzman, Yuhki Nakatake, Carole Stagg, Lioudmila Sharova, Yong Qian, Dawood Dudekula, Sarah Sheer, Jean S Cadet, Tetsuya Hirata, Hsih-Te Yang, Ilya Goldberg, Michele K Evans, Dan L Longo, David Schlessinger, and Minoru S H Ko. "Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells". In: *Scientific Reports* 3.1 (2013), p. 1390.

[263] Fei Sun, Constantinos Chronis, Michael Kronenberg, Xiao-Fen Chen, Trent Su, Fides D Lay, Kathrin Plath, Siavash K Kurdistani, and Michael F Carey. "Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods". In: *Molecular Cell* 73.2 (2019), 250–263.e5.

[264] Michelle Percharde, Fabrice Lavial, J.-H. Ng, Vibhor Kumar, Rute A Tomaz, Nadine Martin, J.-C. Yeo, Jesús Gil, Shyam Prabhakar, H.-H. Ng, Malcolm G Parker, and Véronique Azuara. "Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming". In: *Genes & Development* 26.20 (2012), pp. 2286–2298.

[265] Hitoshi Niwa, Kazuya Ogawa, Daisuke Shimosato, and Kenjiro Adachi. "A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells". In: *Nature* 460.7251 (2009), pp. 118–122.

[266] Ronan Russell, Marcus Ilg, Qiong Lin, Guangming Wu, André Lechel, Wendy Bergmann, Tim Eiseler, Leonhard Linta, Pavan P. Kumar, Moritz Klingenstein, Kenjiro Adachi, Meike Hohwieler, Olena Sakk, Stefanie Raab, Anne Moon, Martin Zenke, Thomas Seufferlein, Hans R. Schöler, Anett Illing, Stefan Liebau, and Alexander Kleger. "A Dynamic Role of TBX3 in the Pluripotency Circuitry". In: *Stem Cell Reports* 5.6 (2015), pp. 1155–1170.

[267] Rui Lu, Acong Yang, and Ying Jin. "Dual functions of T-box 3 (Tbx3) in the control of self-renewal and extraembryonic endoderm differentiation in mouse embryonic stem cells". In: *Journal of Biological Chemistry* 286.10 (2011), pp. 8425–8436.

[268] Apriliana E R Kartikasari, Josie X Zhou, Murtaza S Kanji, David N Chan, Arjun Sinha, Anne Grapin-Botton, Mark A Magnuson, William E Lowry, and Anil Bhushan. "The histone demethylase Jmjd3 sequentially associates with the transcription factors Tbx3 and Eomes to drive endoderm differentiation." In: *The EMBO journal* 32.10 (2013), pp. 1393–408.

[269] Avinash Waghray, Néstor Saiz, Anitha D. Jayaprakash, Ana G. Freire, Dmitri Papatsenko, Carlos Filipe Pereira, Dung Fang Lee, Ran Brosh, Betty Chang, Henia Darr, Julian Gingold, Kevin Kelley, Christoph Schaniel, Anna Katerina Hadjantonakis, and Ihor R. Lemischka. "Tbx3 Controls Dppa3 Levels and Exit from Pluripotency toward Mesoderm". In: *Stem Cell Reports* 5.1 (2015), pp. 97–110.

[270] Miguel Fidalgo, Xin Huang, Diana Guallar, Carlos Sanchez-Priego, Victor Julian Valdes, Arven Saunders, Junjun Ding, Wen Shu Wu, Carlos Clavel, and Jianlong Wang. "Zfp281 Coordinates Opposing Functions of Tet1 and Tet2 in Pluripotent States". In: *Cell Stem Cell* 19.3 (2016), pp. 355–369.

[271] Xin Huang, Sophie Balmer, Fan Yang, Miguel Fidalgo, Dan Li, Diana Guallar, Anna Katerina Hadjantonakis, and Jianlong Wang. "Zfp281 is essential for mouse epiblast maturation through transcriptional and epigenetic control of Nodal signaling". In: *eLife* 6 (2017).

[272] Miguel Fidalgo, Francesco Faiola, C.-F. Pereira, Junjun Ding, Arven Saunders, Julian Gingold, Christoph Schaniel, Ihor R. Lemischka, J. C. R. Silva, and Jianlong Wang. "Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming". In: *Proceedings of the National Academy of Sciences* 109.40 (2012), pp. 16202–16207.

[273] Tilo Kunath, Marc K. Saba-El-Leil, Marwa Almousailleakh, Jason Wray, Sylvain Meloche, and Austin Smith. "FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment". In: *Development* 134.16 (2007), pp. 2895–2902.

[274] Qi Long Ying, Jason Wray, Jennifer Nichols, Laura Batlle-Morera, Bradley Doble, James Woodgett, Philip Cohen, and Austin Smith. "The ground state of embryonic stem cell self-renewal". In: *Nature* 453.7194 (2008), pp. 519–523.

[275] Hendrik Marks, Tüzer Kalkan, Roberta Menafra, Sergey Denissov, Kenneth Jones, Helmut Hofemeister, Jennifer Nichols, Andrea Kranz, A. Francis Stewart, Austin Smith, and Hendrik G. Stunnenberg. "The Transcriptional and Epigenomic Foundations of Ground State Pluripotency". In: *Cell* 149.3 (2012), pp. 590–604.

[276] Gabriella Ficz, Timothy A. Hore, Fátima Santos, Heather J. Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, Simon J. Cook, Simon Andrews, Miguel R. Branco, and Wolf Reik. "FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency". In: *Cell Stem Cell* 13.3 (2013), pp. 351–359.

[277] Ehsan Habibi, Arie B. Brinkman, Julia Arand, Leonie I. Kroeze, Hindrik H.D. Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C. Hubner, Rosaria Benedetti, Lucia Altucci, Joop H. Jansen, Jörn Walter, Ivo G. Gut, Hendrik Marks, and Hendrik G. Stunnenberg. "Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells". In: *Cell Stem Cell* 13.3 (2013), pp. 360–369.

[278] Silvia Domcke, Anaïs Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. "Competition between DNA methylation and transcription factors determines binding of NRF1". In: *Nature* 528.7583 (2015), pp. 575–579.

[279] Xiaofei Zhang, Arne H Smits, Gabrielle Ba van Tilburg, Huib Ovaa, Wolfgang Huber, and Michiel Vermeulen. "Proteome-wide identification of ubiquitin interactions using UbIA-MS". In: *Nature Protocols* 13.3 (2018), pp. 530–550.

[280] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.Suppl 1 (2002), S96–S104.

[281] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47–e47.

[282] Wang Xi and Michael A. Beer. "Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy". In: *PLOS Computational Biology* 14.12 (2018), e1006625.

[283] Masashi Yamaji, Jun Ueda, Katsuhiko Hayashi, Hiroshi Ohta, Yukihiro Yabuta, Kazuki Kurimoto, Ryuichiro Nakato, Yasuhiro Yamada, Katsuhiko Shirahige, and Mitinori Saitou. "PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells". In: *Cell Stem Cell* 12.3 (2013), pp. 368–382.

[284] Ye Ji Sim, Min Seong Kim, Abeer Nayfeh, Ye Jin Yun, Su Jin Kim, Kyung Tae Park, Chang Hoon Kim, and Kye Seong Kim. "2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms". In: *Stem Cell Reports* 8.5 (2017), pp. 1312–1328.

[285] Dae-Kwan Kim, Young Cha, Hee-Jin Ahn, Gwangil Kim, and Kyung-Soon Park. "Lefty1 and Lefty2 Control the Balance Between Self-Renewal and Pluripotent Differentiation of Mouse Embryonic Stem Cells". In: *Stem Cells and Development* 23.5 (2014), pp. 457–466.

[286] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B. Vega, Eleanor Wong, Yuriy L. Orlov, Weiwei Zhang, Jianming Jiang, Yuin Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing Kin Sung, Neil D. Clarke, Chia Lin Wei, and Huck Hui Ng. "Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells". In: *Cell* 133.6 (2008), pp. 1106–1117.

[287] David Brocks, Christopher R. Schmidt, Michael Daskalakis, Hyo Sik Jang, Nakul M. Shah, Daofeng Li, Jing Li, Bo Zhang, Yiran Hou, Sara Laudato, Daniel B. Lipka, Johanna Schott, Holger Bierhoff, Yassen Assenov, Monika Helf, Alzbeta Ressnerova, Md Saiful Islam, Anders M. Lindroth, Simon Haas, Marieke Essers, Charles D. Imbusch, Benedikt Brors, Ina Oehme, Olaf Witt, Michael Lübbert, Jan Philipp Mallm, Karsten Rippe, Rainer Will, Dieter Weichenhan, Georg Stoecklin, Clarissa Gerhäuser, Christopher C. Oakes, Ting Wang, and Christoph Plass. "DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats". In: *Nature Genetics* 49.7 (2017), pp. 1052–1060.

[288] Feng Yue, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D. Pope, Yin Shen, Dmitri D. Pervouchine, Sarah Djebali, Robert E. Thurman, Rajinder Kaul, Eric Rynes, Anthony Kirilusha, Georgi K. Marinov, Brian A. Williams, Diane Trout, Henry Amrhein, Katherine Fisher-Aylor, Igor Antoshechkin, Gilberto DeSalvo, Lei Hoon See, Meagan Fastuca, Jorg Drenkow, Chris Zaleski, Alex Dobin, Pablo Prieto, Julien Lagarde, Giovanni Bussotti, Andrea Tanzer, Olgert Denas, Kanwei Li, M. A. Bender, Miaohua Zhang, Rachel Byron, Mark T. Groudine, David McCleary, Long Pham, Zhen Ye, Samantha Kuan, Lee Edsall, Yi Chieh Wu, Matthew D. Rasmussen, Mukul S. Bansal, Manolis Kellis, Cheryl A. Keller, Christapher S. Morrissey, Tejaswini

Mishra, Deepti Jain, Nergiz Dogan, Robert S. Harris, Philip Cayting, Trupti Kawli, Alan P. Boyle, Ghia Euskirchen, Anshul Kundaje, Shin Lin, Yiing Lin, Camden Jansen, Venkat S. Malladi, Melissa S. Cline, Drew T. Erickson, Vanessa M. Kirkup, Katrina Learned, Cricket A. Sloan, Kate R. Rosenbloom, Beatriz Lacerda De Sousa, Kathryn Beal, Miguel Pignatelli, Paul Flicek, Jin Lian, Tamer Kahveci, Dongwon Lee, W. James Kent, Miguel Ramalho Santos, Javier Herrero, Cedric Notredame, Audra Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Theresa Canfield, Peter J. Sabo, Matthew S. Wilken, Thomas A. Reh, Erika Giste, Anthony Shafer, Tanya Kutyavin, Eric Haugen, Douglas Dunn, Alex P. Reynolds, Shane Neph, Richard Humbert, R. Scott Hansen, Marella De Bruijn, Licia Selleri, Alexander Rudensky, Steven Josefowicz, Robert Samstein, Evan E. Eichler, Stuart H. Orkin, Dana Levasseur, Thalia Papayannopoulou, Kai Hsin Chang, Arthur Skoultchi, Srikanta Gosh, Christine Disteche, Piper Treuting, Yanli Wang, Mitchell J. Weiss, Gerd A. Blobel, Xiaoyi Cao, Sheng Zhong, Ting Wang, Peter J. Good, Rebecca F. Lowdon, Leslie B. Adams, Xiao Qiao Zhou, Michael J. Pazin, Elise A. Feingold, Barbara Wold, James Taylor, Ali Mortazavi, Sherman M. Weissman, John A. Stamatoyannopoulos, Michael P. Snyder, Roderic Guigo, Thomas R. Gingeras, David M. Gilbert, Ross C. Hardison, Michael A. Beer, and Bing Ren. “A comparative encyclopedia of DNA elements in the mouse genome”. In: *Nature* 515.7527 (2014), pp. 355–364.

[289] Galih Kunarso, Na Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun Shen Chan, Huck Hui Ng, and Guillaume Bourque. “Transposable elements have rewired the core regulatory network of human embryonic stem cells”. In: *Nature Genetics* 42.7 (2010), pp. 631–634.

[290] Xiao Ou Zhang, Thomas R. Gingeras, and Zhiping Weng. “Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function”. In: *Genome Research* 29.9 (2019), pp. 1402–1414.

[291] Christopher D. Todd, Özgen Deniz, Darren Taylor, and Miguel R. Branco. “Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells”. In: *eLife* 8 (2019).

[292] Yaqiang Cao, Guoyu Chen, Gang Wu, Xiaoli Zhang, Joseph McDermott, Xingwei Chen, Chi Xu, Quanlong Jiang, Zhaoxiong Chen, Yingying Zeng, Daosheng Ai, Yi Huang, and Jing Dong J. Han. “Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions”. In: *Genome Research* 29.1 (2019), pp. 40–52.

[293] Christina Galonska, Zachary D. Smith, and Alexander Meissner. “In vivo and in vitro dynamics of undifferentiated embryonic cell transcription factor 1”. In: *Stem Cell Reports* 2.3 (2014), pp. 245–252.

[294] David Landeira, Hakan Bagci, Andrzej R. Malinowski, Karen E. Brown, Jorge Soza-Ried, Amelie Feytout, Zoe Webster, Elodie Ndjetehe, Irene Cantone, Helena G. Asenjo, Neil Brockdorff, Thomas Carroll, Matthias Merkenschlager, and Amanda G. Fisher. “Jarid2 coordinates Nanog expression and PCP/Wnt signaling required for efficient ESC differentiation and early embryo development”. In: *Cell Reports* 12.4 (2015), pp. 573–586.

[295] Shen-Hsi Yang, Munazah Andrabi, Rebecca Biss, Syed Murtuza Baker, Mudassar Iqbal, and Andrew D. Sharrocks. "ZIC3 Controls the Transition from Naive to Primed Pluripotency". In: *Cell Reports* 27.11 (2019), 3215–3227.e6.

[296] Nicola Festuccia, Nick Owens, and Pablo Navarro. "Esrrb, an estrogen-related receptor involved in early development, pluripotency, and reprogramming". In: *FEBS Letters* 592.6 (2018), pp. 852–877.

[297] Linda Shushan Lim, Yuin-Han Loh, Weiwei Zhang, Yixun Li, Xi Chen, Yinan Wang, Manjiri Bakre, Huck-Hui Ng, and Lawrence W. Stanton. "Zic3 Is Required for Maintenance of Pluripotency in Embryonic Stem Cells". In: *Molecular Biology of the Cell* 18.4 (2007), pp. 1348–1358.

[298] Alexander Marson, Stuart S. Levine, Megan F. Cole, Garrett M. Frampton, Tobias Brambrink, Sarah Johnstone, Matthew G. Guenther, Wendy K. Johnston, Marius Wernig, Jamie Newman, J. Mauro Calabrese, Lucas M. Dennis, Thomas L. Volkert, Sumeet Gupta, Jennifer Love, Nancy Hannett, Phillip A. Sharp, David P. Bartel, Rudolf Jaenisch, and Richard A. Young. "Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells". In: *Cell* 134.3 (2008), pp. 521–533.

[299] Linda Shushan Lim, Felicia Huimei Hong, Galih Kunarso, and Lawrence W. Stanton. "The Pluripotency Regulator Zic3 Is a Direct Activator of the Nanog Promoter in ESCs". In: *STEM CELLS* 28.11 (2010), pp. 1961–1969.

[300] Hyun Woo Choi, Jin Young Joo, Yean Ju Hong, Jong Soo Kim, Hyuk Song, Jeong Woong Lee, Guangming Wu, Hans R. Schöler, and Jeong Tae Do. "Distinct Enhancer Activity of Oct4 in Naive and Primed Mouse Pluripotency". In: *Stem Cell Reports* 7.5 (2016), pp. 911–926.

[301] Xiaoxiao Zhang, Kevin A. Peterson, X. Shirley Liu, Andrew P. McMahon, and Shinsuke Ohba. "Gene Regulatory Networks Mediating Canonical Wnt Signal-Directed Control of Pluripotency and Differentiation in Embryo Stem Cells". In: *STEM CELLS* 31.12 (2013), pp. 2667–2679.

[302] Avanti Shrikumar, Eva Prakash, and Anshul Kundaje. "GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs". In: *Bioinformatics* 35.14 (2019), pp. i173–i182.

[303] Megan F. Cole, Sarah E. Johnstone, Jamie J. Newman, Michael H. Kagey, and Richard A. Young. "Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells". In: *Genes & Development* 22.6 (2008), pp. 746–755.

[304] Jason Wray, Tüzer Kalkan, Sandra Gomez-Lopez, Dominik Eckardt, Andrew Cook, Rolf Kemler, and Austin Smith. "Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation". In: *Nature Cell Biology* 13.7 (2011), pp. 838–845.

[305] Vicki Metzis, Sebastian Steinhauser, Edvinas Pakanavicius, Mina Gouti, Despina Stamataki, Kenzo Ivanovitch, Thomas Watson, Teresa Rayon, S. Neda Mousavy Gharavy, Robin Lovell-Badge, Nicholas M. Luscombe, and James Briscoe. "Nervous System Regionalization Entails Axial Allocation before Neural Differentiation". In: *Cell* 175.4 (2018), 1105–1118.e17.

[306] Claudio D. Stern, Jeroen Charite, Jacqueline Deschamps, Denis Duboule, Anthony J. Durston, Marie Kmita, Jean-Francois Nicolas, Isabel Palmeirim, Jim C. Smith, and Lewis Wolpert. "Head-tail patterning of the vertebrate embryo: one, two or many unresolved problems?" In: *The International Journal of Developmental Biology* 50.1 (2006), pp. 3–15.

[307] H Spemann and H Mangold. "Induction of embryonic primordia by implantation of organizers from a different species. 1923." In: *The International Journal of Developmental Biology* 45.1 (2001), pp. 13–38.

[308] C D Stern. "Initial patterning of the central nervous system: how many organizers?" In: *Nature reviews. Neuroscience* 2.2 (2001), pp. 92–8.

[309] D. Henrique, E. Abranches, L. Verrier, and K. G. Storey. "Neuromesodermal progenitors and the making of the spinal cord". In: *Development* 142.17 (2015), pp. 2864–2875.

[310] Mina Gouti, Anestis Tsakiridis, Filip J. Wymeersch, Yali Huang, Jens Kleinjung, Valerie Wilson, and James Briscoe. "In Vitro Generation of Neuromesodermal Progenitors Reveals Distinct Roles for Wnt Signalling in the Specification of Spinal Cord and Paraxial Mesoderm Identity". In: *PLOS Biology* 12.8 (2014), e1001937.

[311] Mina Gouti, Julien Delile, Despina Stamataki, Filip J Wymeersch, Yali Huang, Jens Kleinjung, Valerie Wilson, and James Briscoe. "A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development". In: *Developmental Cell* 41.3 (2017), 243–261.e7.

[312] Shinji Takada, Kevin L. Stark, Martin J. Shea, Galya Vassileva, J A McMahon, and Andrew P. McMahon. "Wnt-3a regulates somite and tailbud formation in the mouse embryo." In: *Genes & Development* 8.2 (1994), pp. 174–189.

[313] Terry P. Yamaguchi, Shinji Takada, Yoshiaki Yoshikawa, Nongying Wu, and Andrew P. McMahon. "T (Brachyury) is a direct target of Wnt3a during paraxial mesoderm specification". In: *Genes & Development* 13.24 (1999), pp. 3185–3190.

[314] Teddy Young, Jennifer Elizabeth Rowland, Cesca van de Ven, Monika Bialecka, Ana Novoa, Marta Carapuco, Johan van Nes, Wim de Graaff, Isabelle Duluc, Jean Noël Freund, Felix Beck, Moises Mallo, and Jacqueline Deschamps. "Cdx and Hox Genes Differentially Regulate Posterior Axial Growth in Mammalian Embryos". In: *Developmental Cell* 17.4 (2009), pp. 516–526.

[315] Shilu Amin, Roel Neijts, Salvatore Simmini, Carina van Rooijen, Sander C Tan, Lennart Kester, Alexander van Oudenaarden, Menno P Creyghton, and Jacqueline Deschamps. "Cdx and T Brachyury Co-activate Growth Signaling in the Embryonic Axial Progenitor Niche". In: *Cell Reports* 17.12 (2016), pp. 3165–3177.

[316] Anestis Tsakiridis, Yali Huang, Guillaume Blin, Stavroula Skylaki, Filip Wymeersch, Rodrigo Osorno, Costas Economou, Eleni Karagianni, Suling Zhao, Sally Lowell, and Valerie Wilson. "Distinct Wnt-driven primitive streak-like populations reflect in vivo lineage precursors". In: *Development* 141.6 (2014), pp. 1209–1221.

[317] David A. Turner, Penelope C. Hayward, Peter Baillie-Johnson, P. Rue, Rebecca Broome, Fernando Faunes, and A. Martinez Arias. "Wnt/ -catenin and FGF signalling direct the specification and maintenance of a neuromesodermal axial progenitor in ensembles of mouse embryonic stem cells". In: *Development* 141.22 (2014), pp. 4243–4253.

[318] Frederic Koch, Manuela Scholze, Lars Wittler, Dennis Schifferl, Smita Sudheer, Phillip Grote, Bernd Timmermann, Karol Macura, and Bernhard G Herrmann. "Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors". In: *Developmental Cell* 42.5 (2017), 514–526.e7.

[319] Claire S. Simon and Anna-Katerina Hadjantonakis. "Top to Tail: Anterior-Posterior Patterning Precedes Regional Nervous System Identity". In: *Cell* 175.4 (2018), pp. 905–907.

[320] Ryan Rickels and Ali Shilatifard. "Enhancer Logic and Mechanics in Development and Disease". In: *Trends in Cell Biology* 28.8 (2018), pp. 608–630.

[321] Masanori Uchikawa, Yoshiko Ishida, Tatsuya Takemoto, Yusuke Kamachi, and Hisato Kondoh. "Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals". In: *Developmental Cell* 4.4 (2003), pp. 509–519.

[322] Yusuke Kamachi and Hisato Kondoh. "Sox proteins: regulators of cell fate specification and differentiation". In: *Development* 140.20 (2013), pp. 4129–4144.

[323] Michael A. Lodato, Christopher W. Ng, Joseph A. Wamstad, Albert W. Cheng, Kevin K. Thai, Ernest Fraenkel, Rudolf Jaenisch, and Laurie A. Boyer. "SOX2 Co-Occupies Distal Enhancer Elements with Distinct POU Factors in ESCs and NPCs to Specify Cell State". In: *PLOS Genetics* 9.2 (2013), e1003288.

[324] Hisato Kondoh, Shinji Takada, and Tatsuya Takemoto. "Axial level-dependent molecular and cellular mechanisms underlying the genesis of the embryonic neural plate". In: *Development, Growth & Differentiation* 58.5 (2016), pp. 427–436.

[325] Roel Neijts, Shilu Amin, Carina van Rooijen, Sander Tan, Menno P Creyghton, Wouter de Laat, and Jacqueline Deschamps. "Polarized regulatory landscape and Wnt responsiveness underlie Hox activation in embryos". In: *Genes & Development* 30.17 (2016), pp. 1937–1942.

[326] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414 (2012), pp. 57–74.

[327] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, Kathrina C Onate, Keenan Graham, Stuart R Miyasato, Timothy R Dreszer, J Seth Strattan, Otto Jolanki, Forrest Y Tanaka, and J Michael Cherry. "The Encyclopedia of DNA elements (ENCODE): data portal update". In: *Nucleic Acids Research* 46.D1 (2018), pp. D794–D801.

[328] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1995), pp. 289–300.

[329] Hana Imrichová, Gert Hulselmans, Zeynep Kalender Atak, Delphine Potier, and Stein Aerts. "i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly". In: *Nucleic Acids Research* 43.W1 (2015), W57–W64.

[330] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework". In: *Nucleic Acids Research* 46.D1 (2018), pp. D260–D266.

[331] J. Korhonen, P. Martinmaki, C. Pizzi, P. Rastas, and E. Ukkonen. "MOODS: fast search for position weight matrix matches in DNA sequences". In: *Bioinformatics* 25.23 (2009), pp. 3181–3182.

[332] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data". In: *Genome Research* 21.3 (2011), pp. 447–455.

[333] Esteban O Mazzoni, Shaun Mahony, Mirza Peljto, Tulsi Patel, Seraphim R Thornton, Scott McCuine, Christopher Reeder, Laurie A Boyer, Richard A Young, David K Gifford, and Hynek Wichterle. "Saltatory remodeling of Hox chromatin in response to rostrocaudal patterning signals". In: *Nature Neuroscience* 16.9 (2013), pp. 1191–1198.

[334] Benilton S. Carvalho and Rafael A. Irizarry. "A framework for oligonucleotide microarray preprocessing". In: *Bioinformatics* 26.19 (2010), pp. 2363–2367.

[335] Li Teng, Bing He, Jiahui Wang, and Kai Tan. "4DGenome: a comprehensive database of chromatin interactions". In: *Bioinformatics* 31.15 (2015), pp. 2560–2564.

[336] A Pattyn, X Morin, H Cremer, C Goridis, and J F Brunet. "Expression and interactions of the two closely related homeobox genes Phox2a and Phox2b during neurogenesis." In: *Development* 124.20 (1997), pp. 4065–75.

[337] Marc Friedli, Isabelle Barde, Mélanie Arcangeli, Sonia Verp, Alexandra Quazzola, Jozsef Zakany, Nathalie Lin-Marq, Daniel Robyr, Catia Attanasio, François Spitz, Denis Duboule, Didier Trono, and Stylianos E. Antonarakis. "A Systematic Enhancer Screen Using Lentivector Transgenesis Identifies Conserved and Non-Conserved Functional Elements at the Olig1 and Olig2 Locus". In: *PLOS ONE* 5.12 (2010), e15741.

[338] Sarah K. Denny, Dian Yang, Chen-Hua Chuang, Jennifer J. Brady, Jing Shan Lim, Barbara M. Grüner, Shin-Heng Chiou, Alicia N. Schep, Jessika Baral, Cécile Hamard, Martine Antoine, Marie Wislez, Christina S. Kong, Andrew J. Connolly, Kwon-Sik Park, Julien Sage, William J. Greenleaf, Monte M. Winslow, A. Alevizopoulos, Y. Dusserre, M. Tsai-Pflugfelder, T. von der Weid, W. Wahli, N. Mermod, A. Augustyn, M. Borromeo, T. Wang, J. Fujimoto, C. Shao, P.D. Dospoy, V. Lee, C. Tan, J.P. Sullivan, J.E. Larsen, Et Al., J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C.J. Wilson, J. Lehár, G.V. Kryukov, D. Sonkin, Et Al., O. Bell, V.K. Tiwari, N.H. Thomä, D. Schübeler, P.V. Benos, M.L. Bulyk, G.D. Stormo, J. Betancourt, S. Katzman, B. Chen, P. Blomquist, Q. Li, O. Wrange, J.D. Buenrostro, P.G. Giresi, L.C. Zaba, H.Y. Chang, W.J. Greenleaf, M. Burger, A. Glodek, T. Hartmann, A. Schmitt-Gräff, L.E. Silberstein, N. Fujii, T.J. Kipps, J.A. Burger, C.-Y. Chang, H.A. Pasolli, E.G. Giannopoulou, G. Guasch, R.M. Gronostajski, O.

Elemento, E. Fuchs, J. Chen, R.B. Deal, J.G. Henikoff, S. Henikoff, A.L. Dooley, M.M. Winslow, D.Y. Chiang, S. Banerji, N. Stransky, T.L. Dayton, E.L. Snyder, S. Senna, C.A. Whittaker, R.T. Bronson, Et Al., K. Foley, A.A. Rucki, Q. Xiao, D. Zhou, A. Leubner, G. Mo, J. Kleponis, A.A. Wu, R. Sharma, Q. Jiang, Et Al., J. George, J.S. Lim, S.J. Jang, Y. Cun, L. Ozretić, G. Kong, F. Leenders, X. Lu, L. Fernández-Cuesta, G. Bosco, Et Al., M.M. Grabowska, A.D. Elliott, D.J. DeGraff, P.D. Anderson, G. Anumanthan, H. Yamashita, Q. Sun, D.B. Friedman, D.L. Hachey, X. Yu, Et Al., R.M. Gronostajski, L. Harris, L.A. Genovesi, R.M. Gronostajski, B.J. Wainwright, M. Piper, P.B. Hebbar, T.K. Archer, J.M. Hou, M.G. Krebs, L. Lancashire, R. Sloane, A. Backen, R.K. Swain, L.J. Priest, A. Greystoke, C. Zhou, K. Morris, Et Al., G.P. Kalemkerian, W. Akerley, P. Bogner, H. Borghaei, L.Q. Chow, R.J. Downey, L. Gandhi, A.K. Ganti, R. Govindan, J.C. Grecula, National Comprehensive Cancer Network, Et Al., S.Y. Ko, G.L. Blatch, C.R. Dass, C.Y. Lin, J. Loven, P.B. Rahl, R.M. Paranal, C.B. Burge, J.E. Bradner, T.I. Lee, R.A. Young, M.I. Love, W. Huber, S. Anders, M. Meisterernst, I. Gander, L. Rogge, E.L. Winnacker, R. Meuwissen, S.C. Linn, R.I. Linnoila, J. Zevenhoven, W.J. Mooi, A. Berns, H.P. Mohammad, K.N. Smitheman, C.D. Kamat, D. Soong, K.E. Federowicz, G.S. Van Aller, J.L. Schneck, J.D. Carson, Y. Liu, M. Butticello, Et Al., M.D. Muzumdar, B. Tasic, K. Miyamichi, L. Li, L. Luo, K. Nakazawa, K. Kurishima, T. Tamura, K. Kagohashi, H. Ishikawa, H. Satoh, N. Hizawa, J.K. Osborne, J.E. Larsen, M.D. Shields, J.X. Gonzales, D.S. Shames, M. Sato, A. Kulkarni, I.I. Wistuba, L. Girard, J.D. Minna, M.H. Cobb, M. Peifer, L. Fernández-Cuesta, M.L. Sos, J. George, D. Seidel, L.H. Kasper, D. Plenker, F. Leenders, R. Sun, T. Zander, Et Al., C.M. Rudin, S. Durinck, E.W. Stawiski, J.T. Poirier, Z. Modrusan, D.S. Shames, E.A. Bergbower, Y. Guan, J. Shin, J. Guillory, Et Al., B.E. Schaffer, K.S. Park, G. Yiu, J.F. Conklin, C. Lin, D.L. Burkhart, A.N. Karnezis, E.A. Sweet-Cordero, J. Sage, A.N. Schep, J.D. Buenrostro, S.K. Denny, K. Schwartz, G. Sherlock, W.J. Greenleaf, D.E. Schones, K. Zhao, N. Sethi, Y. Kang, J.M. Simon, K.E. Hacker, D. Singh, A.R. Brannon, J.S. Parker, M. Weiser, T.H. Ho, P.F. Kuan, E. Jonasch, T.S. Furey, Et Al., G. Steele-Perkins, C. Plachez, K.G. Butz, G. Yang, C.J. Bachurski, S.L. Kinsman, E.D. Litwack, L.J. Richards, R.M. Gronostajski, A.B. Stergachis, S. Neph, A. Reynolds, R. Humbert, B. Miller, S.L. Paige, B. Vernot, J.B. Cheng, R.E. Thurman, R. Sandstrom, Et Al., M.-H. Sung, M.J. Guertin, S. Baek, G.L. Hager, M.-H. Sung, S. Baek, G.L. Hager, R.E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M.T. Maurano, E. Haugen, N.C. Sheffield, A.B. Stergachis, H. Wang, B. Vernot, Et Al., J.P. van Meerbeeck, D.A. Fennell, D.K.M. De Ruysscher, S. Vanharanta, W. Shu, F. Brenet, A.A. Hakimi, A. Heguy, A. Viale, V.E. Reuter, J.J. Hsieh, J.M. Scandura, J. Massagué, J. Vierstra, E. Rynes, R. Sandstrom, M. Zhang, T. Canfield, R.S. Hansen, S. Stehling-Sun, P.J. Sabo, R. Byron, R. Humbert, Et Al., T.C. Voss, G.L. Hager, J. Yang, R.A. Weinberg, K.S. Zaret, S.E. Mango, Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, W.-J. Zhou, Z.H. Geng, S. Chi, W. Zhang, X.F. Niu, S.J. Lan, L. Ma, X. Yang, L.J. Wang, Y.Q. Ding, J.G. Geng, J. Zhu, M. Adli, J.Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P.L. De Jager, and Et Al. "Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility". In: *Cell* 166.2 (2016), pp. 328–342.

[339] Kevin A. Peterson, Yuichi Nishi, Wenxiu Ma, Anastasia Vedenko, Leila Shokri, Xiaoxiao Zhang, Matthew McFarlane, J.-M. Baizabal, Jan Philipp Junker, Alexander van Oudenaarden, Tarjei Mikkelsen, Bradley E. Bernstein, Timothy L. Bailey, Martha

L. Bulyk, Wing H. Wong, and Andrew P. McMahon. “Neural-specific Sox2 input and differential Gli-binding affinity provide context and positional information in Shh-directed neural patterning”. In: *Genes & Development* 26.24 (2012), pp. 2802–2816.

[340] Qi Long Ying, Marios Stavridis, Dean Griffiths, Meng Li, and Austin Smith. “Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture”. In: *Nature Biotechnology* 21.2 (2003), pp. 183–186.

[341] D J Epstein, A P McMahon, and A L Joyner. “Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms.” In: *Development* 126.2 (1999), pp. 281–92.

[342] Omar Abdel Samad. “Integration of anteroposterior and dorsoventral regulation of Phox2b transcription in cranial motoneuron progenitors by homeodomain proteins”. In: *Development* 131.16 (2004), pp. 4071–4083.

[343] Heinz-Georg Belting, Cooduvalli S. Shashikant, and Frank H. Ruddle. “Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology”. In: *Proceedings of the National Academy of Sciences* 95.5 (1998), pp. 2355–2360.

[344] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio. “VISTA Enhancer Browser–a database of tissue-specific human enhancers”. In: *Nucleic Acids Research* 35.Database (2007), pp. D88–D92.

[345] Feng Yue, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D. Pope, Yin Shen, Dmitri D. Pervouchine, Sarah Djebali, Robert E. Thurman, Rajinder Kaul, Eric Rynes, Anthony Kirilusha, Georgi K. Marinov, Brian A. Williams, Diane Trout, Henry Amrhein, Katherine Fisher-Aylor, Igor Antoshechkin, Gilberto DeSalvo, Lei-Hoon See, Meagan Fastuca, Jorg Drenkow, Chris Zaleski, Alex Dobin, Pablo Prieto, Julien Lagarde, Giovanni Bussotti, Andrea Tanzer, Olgert Denas, Kanwei Li, M. A. Bender, Miaohua Zhang, Rachel Byron, Mark T. Groudine, David McCleary, Long Pham, Zhen Ye, Samantha Kuan, Lee Edsall, Yi-Chieh Wu, Matthew D. Rasmussen, Mukul S. Bansal, Manolis Kellis, Cheryl A. Keller, Christapher S. Morrissey, Tejaswini Mishra, Deepti Jain, Nergiz Dogan, Robert S. Harris, Philip Cayting, Trupti Kawli, Alan P. Boyle, Ghia Euskirchen, Anshul Kundaje, Shin Lin, Yiing Lin, Camden Jansen, Venkat S. Malladi, Melissa S. Cline, Drew T. Erickson, Vanessa M. Kirkup, Katrina Learned, Cricket A. Sloan, Kate R. Rosenbloom, Beatriz Lacerda de Sousa, Kathryn Beal, Miguel Pignatelli, Paul Flicek, Jin Lian, Tamer Kahveci, Dongwon Lee, W. James Kent, Miguel Ramalho Santos, Javier Herrero, Cedric Notredame, Audra Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Theresa Canfield, Peter J. Sabo, Matthew S. Wilken, Thomas A. Reh, Erika Giste, Anthony Shafer, Tanya Kutyavin, Eric Haugen, Douglas Dunn, Alex P. Reynolds, Shane Neph, Richard Humbert, R. Scott Hansen, Marella De Bruijn, Licia Selleri, Alexander Rudensky, Steven Josefowicz, Robert Samstein, Evan E. Eichler, Stuart H. Orkin, Dana Levasseur, Thalia Papayannopoulou, Kai-Hsin Chang, Arthur Skoultchi, Srikanta Gosh, Christine Disteche, Piper Treuting, Yanli Wang, Mitchell J. Weiss, Gerd A. Blobel, Xiaoyi Cao, Sheng Zhong, Ting Wang, Peter J. Good, Rebecca F. Lowdon, Leslie B. Adams, Xiao-Qiao Zhou, Michael J. Pazin, Elise A. Feingold,

Barbara Wold, James Taylor, Ali Mortazavi, Sherman M. Weissman, John A. Stamatoyannopoulos, Michael P. Snyder, Roderic Guigo, Thomas R. Gingeras, David M. Gilbert, Ross C. Hardison, Michael A. Beer, and Bing Ren. "A comparative encyclopedia of DNA elements in the mouse genome". In: *Nature* 515.7527 (2014), pp. 355–364.

[346] Julien Delile, Teresa Rayon, Manuela Melchionda, Amelia Edwards, James Briscoe, and Andreas Sagner. "Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord". In: *Development* 146.12 (2019), dev173807.

[347] Yannis E. Mavromatakis, Wei Lin, Emmanouil Metzakopian, Anna L.M. Ferri, Carol H. Yan, Hiroshi Sasaki, Jeff Whisett, and Siew-Lan Ang. "Foxa1 and Foxa2 positively and negatively regulate Shh signalling to specify ventral midbrain progenitor identity". In: *Mechanisms of Development* 128.1-2 (2011), pp. 90–103.

[348] Laura Domínguez, Agustín González, and Nerea Moreno. "Ontogenetic Distribution of the Transcription Factor Nkx2.2 in the Developing Forebrain of Xenopus Laevis". In: *Frontiers in Neuroanatomy* 5.MARCH (2011), p. 7.

[349] Dongchang Xiao, Xiaoning Liu, Min Zhang, Min Zou, Qinqin Deng, Dayu Sun, Xuting Bian, Yulong Cai, Yanan Guo, Shuting Liu, Shengguo Li, Evelyn Shiang, Hongyu Zhong, Lin Cheng, Haiwei Xu, Kangxin Jin, and Mengqing Xiang. "Direct reprogramming of fibroblasts into neural stem cells by single non-neural progenitor transcription factor Ptf1a". In: *Nature Communications* 9.1 (2018), p. 2865.

[350] Bennett G. Novitch, Albert I. Chen, and Thomas M. Jessell. "Coordinate regulation of motor neuron subtype identity and pan-neuronal properties by the bHLH repressor Olig2". In: *Neuron* 31.5 (2001), pp. 773–789.

[351] Alexandre Pattyn. "Complementary roles for Nkx6 and Nkx2 class proteins in the establishment of motoneuron identity in the hindbrain". In: *Development* 130.17 (2003), pp. 4149–4159.

[352] Chris T. Dee, Caroline S. Hirst, Yu Huan Shih, Vineeta B. Tripathi, Roger K. Patient, and Paul J. Scotting. "Sox3 regulates both neural fate and differentiation in the zebrafish ectoderm". In: *Developmental Biology* 320.1 (2008), pp. 289–301.

[353] Kallayanee Chawengsaksophak, Wim de Graaff, Janet Rossant, Jacqueline Deschamps, and Felix Beck. "Cdx2 is essential for axial elongation in mouse development". In: *Proceedings of the National Academy of Sciences* 101.20 (2004), pp. 7641–7645.

[354] Ulrika Nordström, Esther Maier, Thomas M Jessell, and Thomas Edlund. "An Early Role for Wnt Signaling in Specifying Neural Patterns of Cdx and Hox Gene Expression and Motor Neuron Subtype Identity". In: *PLOS Biology* 4.8 (2006), e252.

[355] Mirza Peljto, Jeremy S. Dasen, Esteban O. Mazzoni, Thomas M. Jessell, and Hynek Wichterle. "Functional Diversity of ESC-Derived Motor Neuron Subtypes Revealed through Intraspinal Transplantation". In: *Cell Stem Cell* 7.3 (2010), pp. 355–366.

[356] Esteban O Mazzoni, Shaun Mahony, Michael Closser, Carolyn A Morrison, Stephane Nedelec, Damian J Williams, Disi An, David K Gifford, and Hynek Wichterle. "Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity". In: *Nature Neuroscience* 16.9 (2013), pp. 1219–1227.

[357] Heather B. Wood and Vasso Episkopou. “Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages”. In: *Mechanisms of Development* 86.1-2 (1999), pp. 197–201.

[358] Andreas Sagner and James Briscoe. “Establishing neuronal diversity in the spinal cord: a time and a place”. In: *Development* 146.22 (2019), dev182154.

[359] Luis R. Hernandez-Miranda, Thomas Müller, and Carmen Birchmeier. “The dorsal spinal cord and hindbrain: From developmental mechanisms to functional circuits”. In: *Developmental Biology* 432.1 (2017), pp. 34–42.

[360] Roel Neijts, Shilu Amin, Carina van Rooijen, and Jacqueline Deschamps. “Cdx is crucial for the timing mechanism driving colinear Hox activation and defines a trunk segment in the Hox cluster topology”. In: *Developmental Biology* 422.2 (2017), pp. 146–154.

[361] Gregg Duester. “Retinoic Acid Synthesis and Signaling during Early Organogenesis”. In: *Cell* 134.6 (2008), pp. 921–931.

[362] Hitoshi Niwa, Yayoi Toyooka, Daisuke Shimosato, Dan Strumpf, Kadue Takahashi, Rika Yagi, and Janet Rossant. “Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation”. In: *Cell* 123.5 (2005), pp. 917–929.

[363] Wm Gregory Cox and Ali Hemmati-Brivanlou. “Caudalization of neural fate by tissue recombination and bFGF”. In: *Development* 121.12 (1995), pp. 4349–4358.

[364] Jonas Muhr, Enrique Graziano, Sara Wilson, Thomas M. Jessell, and Thomas Edlund. “Convergent inductive signals specify midbrain, hindbrain, and spinal cord identity in Gastrula stage chick embryos”. In: *Neuron* 23.4 (1999), pp. 689–702.

[365] Sara I. Wilson, Anna Rydström, Tolleiv Trimborn, Karl Willert, Roel Musse, Thomas M. Jessell, and Thomas Edlund. “The status of Wnt signalling regulates neural and epidermal fates in the chick embryo”. In: *Nature* 411.6835 (2001), pp. 325–330.

[366] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation”. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 96. 6. National Academy of Sciences, 1999, pp. 2907–2912.

[367] Francis D. Gibbons and Frederick P. Roth. “Judging the quality of gene expression-based clustering methods using gene annotation”. In: *Genome Research* 12.10 (2002), pp. 1574–1581.

[368] Ferdinand Marlétaz, Ignacio Maeso, Laura Faas, Harry V. Isaacs, and Peter W.H. Holland. “Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution”. In: *BMC Biology* 13.1 (2015), p. 56.

[369] Nina M. Brooke, Jordi Garcia-Fernàndez, and Peter W.H. Holland. “The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster”. In: *Nature* 392.6679 (1998), pp. 920–922.

[370] Marek Mlodzik, Anders Fjose, and Walter J. Gehring. “Isolation of caudal , a Drosophila homeo box-containing gene with maternal expression, whose transcripts form a concentration gradient at the pre-blastoderm stage”. In: *The EMBO Journal* 4.11 (1985), pp. 2961–2969.

[371] John F. Mulley, Chi Hua Chiu, and Peter W.H. Holland. “Breakup of a homeobox cluster after genome duplication in teleosts”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.27 (2006), pp. 10369–10372.

[372] Laura Faas and Harry V. Isaacs. “Overlapping functions of Cdx1, Cdx2, and Cdx4 in the development of the amphibian Xenopus tropicalis”. In: *Developmental Dynamics* 238.4 (2009), pp. 835–852.

[373] Detlev Arendt, Jacob M. Musser, Clare V. H. Baker, Aviv Bergman, Connie Cepko, Douglas H. Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D. Laubichler, and Günter P. Wagner. “The origin and evolution of cell types”. In: *Nature Reviews Genetics* 17.12 (2016), pp. 744–757.

[374] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. “Deep learning at base-resolution reveals motif syntax of the cis-regulatory code”. In: *bioRxiv* (2019), p. 737981.

[375] Phil Arnold, A. Scholer, Mikhail Pachkov, Piotr J Balwierz, H. Jorgensen, Michael B Stadler, Erik van Nimwegen, and D. Schubeler. “Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting”. In: *Genome Research* 23.1 (2013), pp. 60–73.

[376] Pratiksha I Thakore, Anthony M D’Ippolito, Lingyun Song, Alexias Safi, Nishkala K Shivakumar, Ami M Kabadi, Timothy E Reddy, Gregory E Crawford, and Charles A Gersbach. “Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements”. In: *Nature Methods* 12.12 (2015), pp. 1143–1149.

[377] Jason C. Klein, Wei Chen, Molly Gasperini, and Jay Shendure. “Identifying Novel Enhancer Elements with CRISPR-Based Screens”. In: *ACS Chemical Biology* 13.2 (2018), pp. 326–332.

[378] Wen Deng, Jack A. Bates, Hai Wei, Michael D. Bartoschek, Barbara Conradt, and Heinrich Leonhardt. “Tunable light and drug induced depletion of target proteins”. In: *Nature Communications* 11.1 (2020), p. 304.

[379] Nisha Rajagopal, Sharanya Srinivasan, Kameron Kooshesh, Yuchun Guo, Matthew D Edwards, Budhaditya Banerjee, Tahin Syed, Bart J M Emons, David K Gifford, and Richard I Sherwood. “High-throughput mapping of regulatory DNA”. In: *Nature Biotechnology* 34.2 (2016), pp. 167–174.

[380] Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan, and Bing Ren. “A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells”. In: *Nature Methods* 14.6 (2017), pp. 629–635.

[381] Naresh Doni Jayavelu, Ajay Jajodia, Arpit Mishra, and R. David Hawkins. “An atlas of silencer elements for the human and mouse genomes”. In: *bioRxiv* (2018), p. 252304.

[382] Di Huang, Hanna M. Petrykowska, Brendan F. Miller, Laura Elnitski, and Ivan Ovcharenko. “Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression”. In: *Genome Research* 29.4 (2019), pp. 657–667.

[383] Ulrike M. Litzenburger, Jason D. Buenrostro, Beijing Wu, Ying Shen, Nathan C. Sheffield, Arwa Kathiria, William J. Greenleaf, and Howard Y. Chang. "Single-cell epigenomic variability reveals functional cancer heterogeneity". In: *Genome Biology* 18.1 (2017), p. 15.

[384] Hannah A Pliner, Jonathan S Packer, José L McFaline-Figueroa, Darren A Cusanovich, Riza M Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C Adey, Frank J Steemers, Jay Shendure, and Cole Trapnell. "Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data". In: *Molecular Cell* 71.5 (2018), 858–871.e8.

[385] Mahdi Zamanighomi, Zhixiang Lin, Timothy Daley, Xi Chen, Zhana Duren, Alicia Schep, William J. Greenleaf, and Wing Hung Wong. "Unsupervised clustering and epigenetic classification of single cells". In: *Nature Communications* 9.1 (2018).

[386] Erik Engelen, Johannes H. Brandsma, Maaike J. Moen, Luca Signorile, Dick H. W. Dekkers, Jeroen Demmers, Christel E. M. Kockx, Zehila Ozgür, Wilfred F. J. van IJcken, Debbie L. C. van den Berg, and Raymond A. Poot. "Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry". In: *Nature Communications* 6.1 (2015), p. 7155.

[387] Sergei Denissov, Helmut Hofemeister, Hendrik Marks, Andrea Kranz, Giovanni Ciotta, Sukhdeep Singh, Konstantinos Anastassiadis, Hendrik G. Stunnenberg, and A. Francis Stewart. "Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant". In: *Development* 141.3 (2014), pp. 526–537.

[388] Gang Li, Raphael Margueron, Manching Ku, Pierre Chambon, Bradley E. Bernstein, and Danny Reinberg. "Jarid2 and PRC2, partners in regulating gene expression". In: *Genes & Development* 24.4 (2010), pp. 368–380.

[389] Aster H. Juan, Stan Wang, Kyung Dae Ko, Hossein Zare, Pei Fang Tsai, Xuesong Feng, Karinna O. Vivanco, Anthony M. Ascoli, Gustavo Gutierrez-Cruz, Jordan Krebs, Simone Sidoli, Adam L. Knight, Roger A. Pedersen, Benjamin A. Garcia, Rafael Casellas, Jizhong Zou, and Vittorio Sartorelli. "Roles of H3K27me2 and H3K27me3 Examined during Fate Specification of Embryonic Stem Cells". In: *Cell Reports* 17.5 (2016), pp. 1369–1382.

[390] Luigi Aloia, Bruno Di Stefano, Alessandro Sessa, Lluis Morey, Alexandra Santanach, Arantxa Gutierrez, Luca Cozzuto, Salvador Aznar Benitah, Thomas Graf, Vania Broccoli, and Luciano Di Croce. "Zrf1 is required to establish and maintain neural progenitor identity". In: *Genes & Development* 28.2 (2014), pp. 182–197.

[391] Maria Bergsland, D. Ramskold, Cécile Zaouter, Susanne Klum, Rickard Sandberg, and Jonas Muhr. "Sequentially acting Sox transcription factors in neural lineage development". In: *Genes & Development* 25.23 (2011), pp. 2453–2464.

[392] Mark D. Borromeo, David M. Meredith, Diogo S. Castro, Joshua C. Chang, Kuang-Chi Tung, Francois Guillemot, and Jane E. Johnson. "A transcription factor network specifying inhibitory versus excitatory neurons in the dorsal spinal cord". In: *Development* 141.14 (2014), pp. 2803–2812.

[393] Heekyung Jung, Julie Lacombe, Esteban O. Mazzoni, Karel F. Liem, Jonathan Grinstein, Shaun Mahony, Debnath Mukhopadhyay, David K. Gifford, Richard A. Young, Kathryn V. Anderson, Hynek Wichterle, and Jeremy S. Dasen. "Global Control of Motor Neuron Topography Mediated by the Repressive Actions of a Single Hox Gene". In: *Neuron* 67.5 (2010), pp. 781–796.

[394] Dale McAninch and Paul Thomas. "Identification of Highly Conserved Putative Developmental Enhancers Bound by SOX3 in Neural Progenitors Using ChIP-Seq". In: *PLoS ONE* 9.11 (2014), e113361.

[395] Tapan Kumar Mistri, Arun George Devasia, Lee Thean Chu, Wei Ping Ng, Florian Halbritter, Douglas Colby, Ben Martynoga, Simon R Tomlinson, Ian Chambers, Paul Robson, and Thorsten Wohland. "Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells". In: *EMBO reports* 16.9 (2015), pp. 1177–1191.

[396] Yuichi Nishi, Xiaoxiao Zhang, Jieun Jeong, Kevin A. Peterson, Anastasia Vedenko, Martha L. Bulyk, Winston A. Hide, and Andrew P. McMahon. "A direct fate exclusion mechanism by Sonic hedgehog-regulated transcriptional repressors". In: *Development* 142.19 (2015), pp. 3286–3293.

Appendix A

Supplementary Figures

A.1 Chapter2 Supplementary Figures

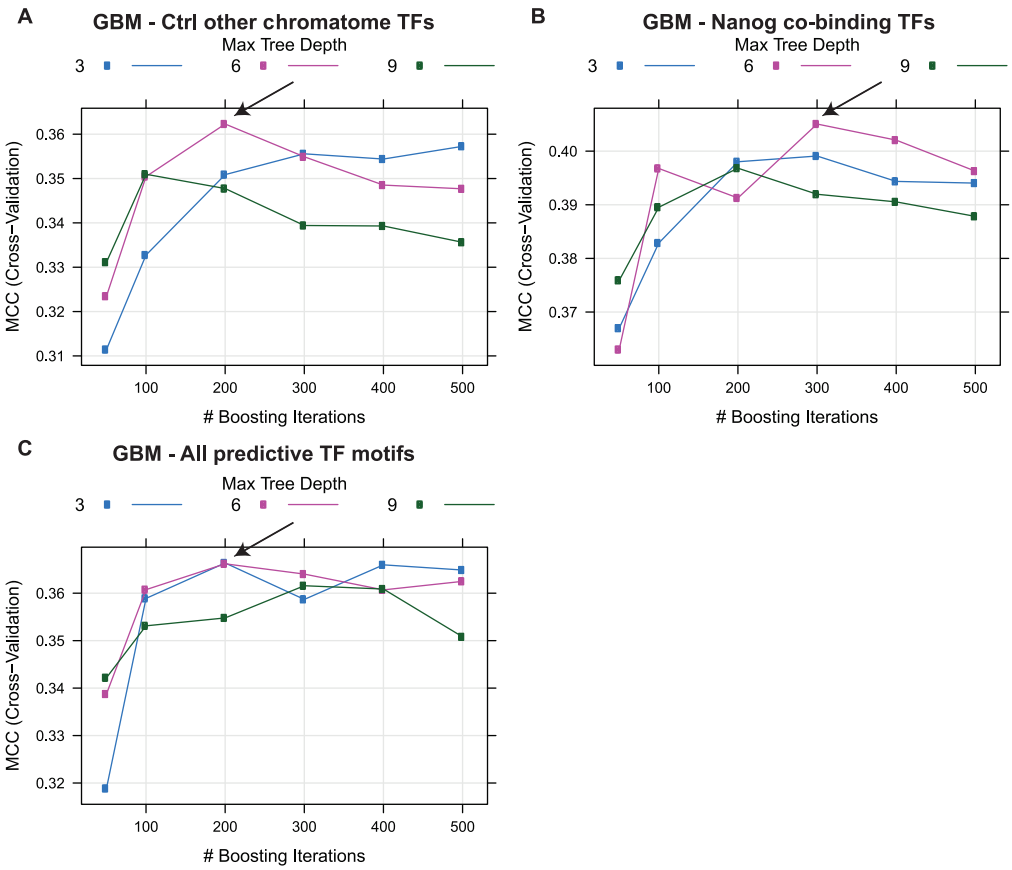


Fig. A.1: Grid-search results from GBM hyperparameter optimisation. Grid-search was performed to optimise 10-fold cross-validated Mathews correlation coefficient (MCC) to find the best values for two hyperparameters: the number of boosting iterations and max tree depth. Results shown for GBMs trained on different feature sets including **(A)** control TFs, **(B)** Nanog co-binding TFs and **(C)** all predictive TFs. Best set of hyperparameters is highlighted with an arrow.

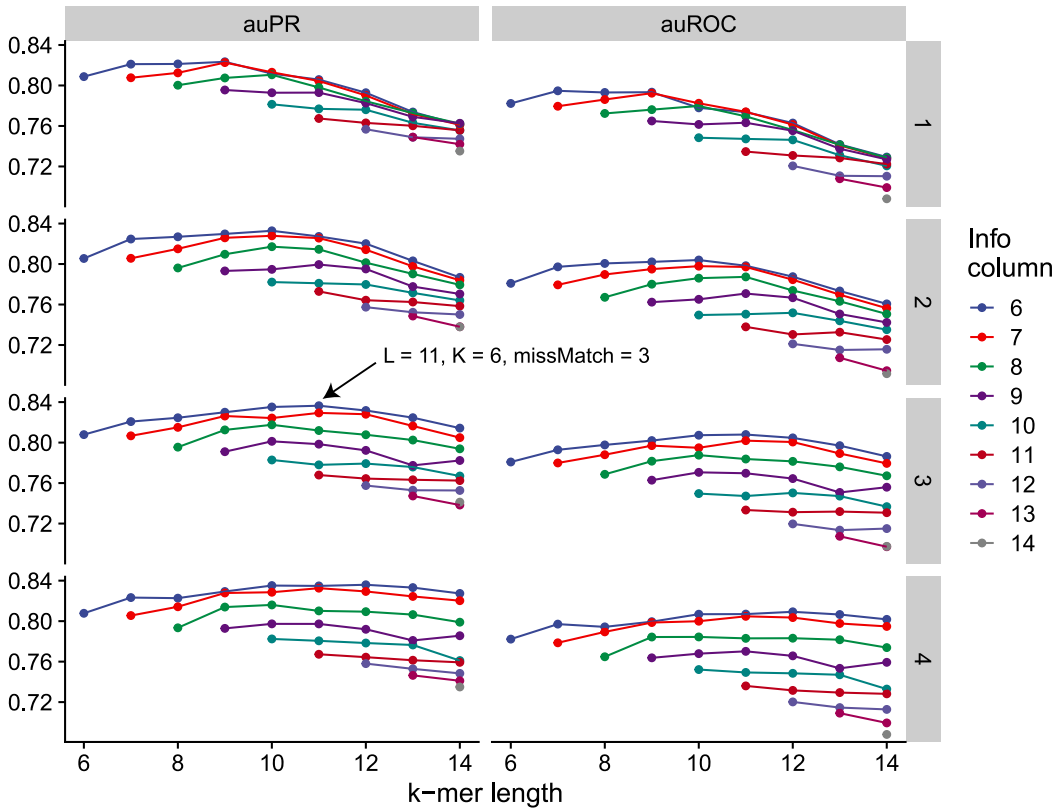


Fig. A.2: Grid-search results for gkm-SVM hyperparameter optimisation. Hyperparameter search results for k-mer lengths (L) between 6-14, number of informative columns (K) and number of miss matches 1-4. 10-fold cross-validated performance measures were computed including Area under precision-recall (PR) curve (*left panel*) and ROC (*right panel*). Best set of hyperparameters is highlighted with an arrow.

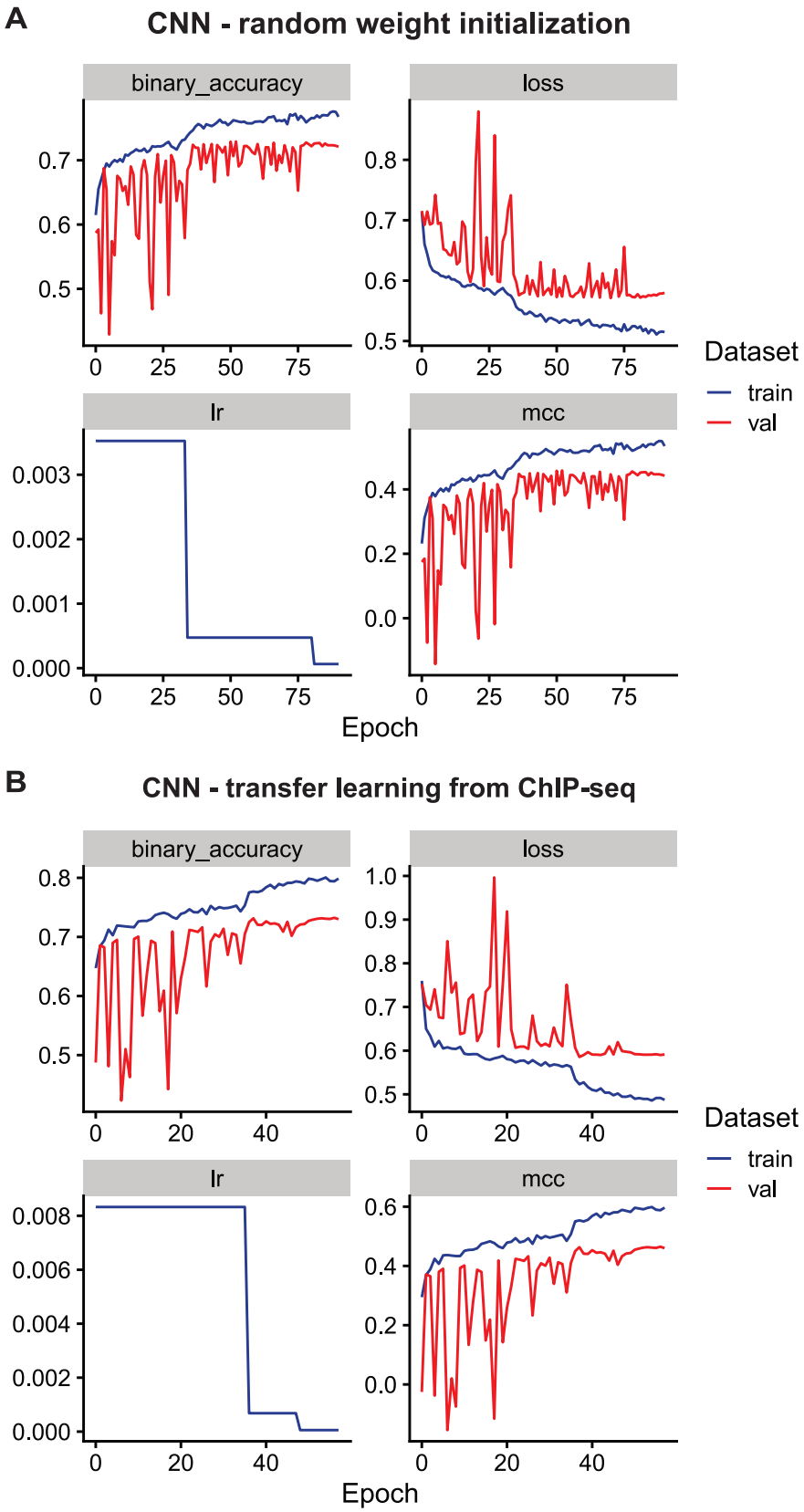


Fig. A.3: Model training history for CNN with best architecture from hyperparameter optimisation. Accuracy, loss, learning rate (lr) and Matthews correlation coefficient (MCC) were tracked for each epoch for training (red) as well as validation (blue) dataset. Training history shown for CNNs initialised with **(A)** random weights and **(B)** transfer learning by initiation with weights from a pre-trained CNN to distinguish Nanog binding from shuffled background.

A.2 Chapter3 Supplementary Figures

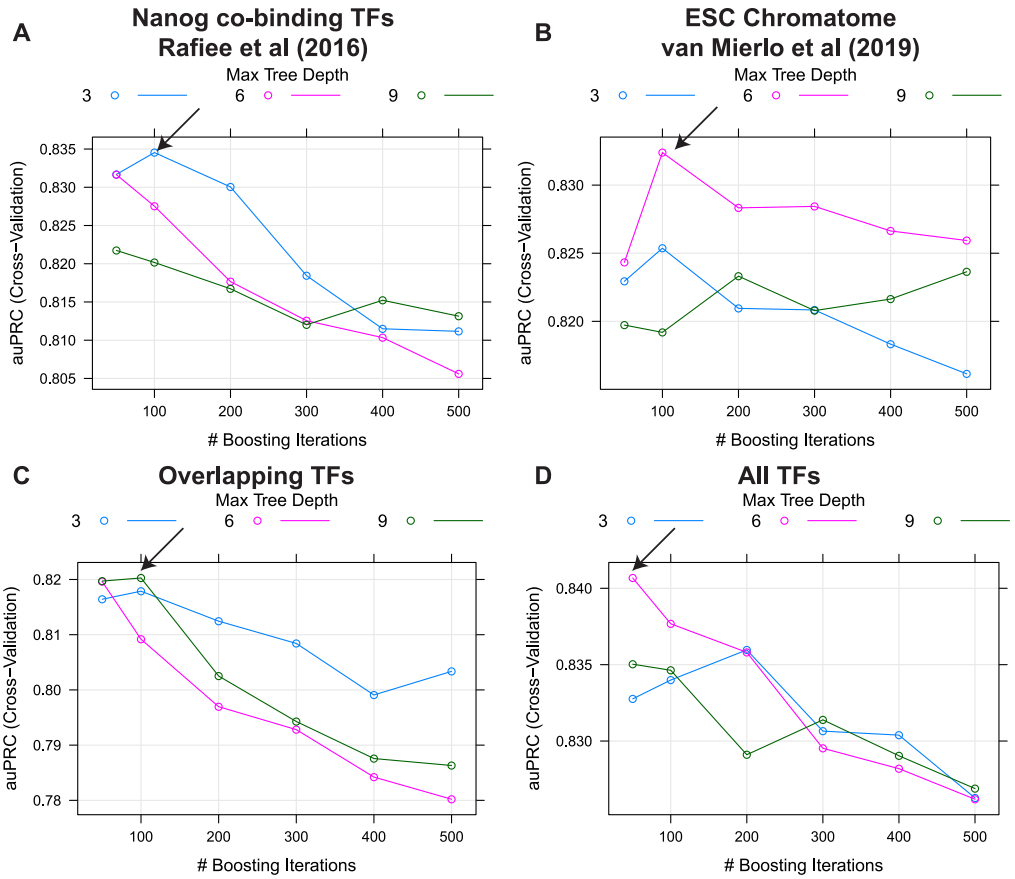


Fig. A.4: Grid-search results from GBM hyperparameter optimisation. Grid-search was performed to optimise 10-fold cross-validated Area under precision-recall (PR) curve to find the best values for two hyperparameters: the number of boosting iterations and max tree depth. Results shown for GBMs trained on different feature sets including **(A)** control TFs, **(B)** Nanog co-binding TFs and **(C)** all predictive TFs. Best set of hyperparameters is highlighted with an arrow.

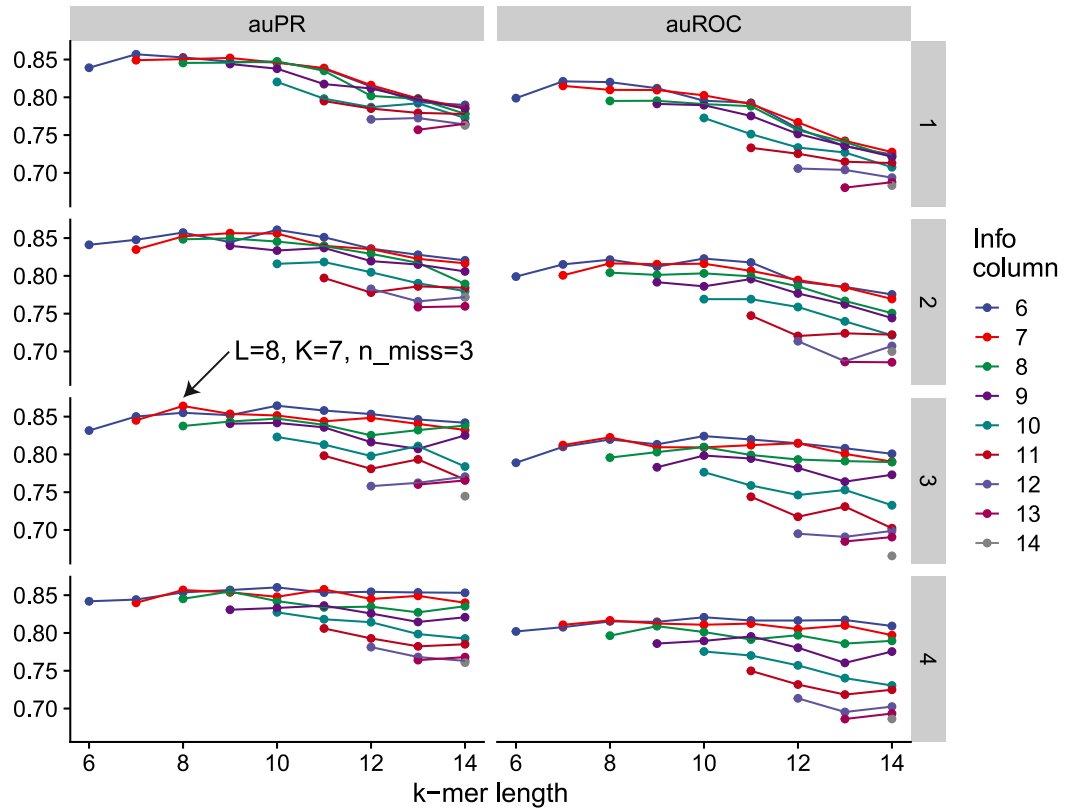


Fig. A.5: Grid-search results for gkm-SVM hyperparameter optimisation. Hyperparameter search results for k-mer lengths (L) between 6-14, number of informative columns (K) and number of miss matches 1-4. 10-fold cross-validated performance measures were computed including Area under precision-recall (PR) curve (*left panel*) and ROC (*right panel*). Best set of hyperparameters is highlighted with an arrow.

Appendix B

Supplementary Tables

B.1 Chapter 2 and 3 Supplementary Tables

Table B.1: ChIP STARR-seq datasets used in Chapter 2 & 3.

Study	Description	Rep	
Steinhauser 2020	ESC 2i Nanog ChIP vector library	4	
Steinhauser 2020	ESC 2i STARR-seq		3
Steinhauser 2020	ESC serum STARR-seq	3	

Table B.2: ChIP-exo datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	ChIP	Rep
Xie 2017 [254]	GSE97304	ESC	Esrrb	1
Xie 2017 [254]	GSE97304	ESC	Stat3	1

Table B.3: CUT&RUN datasets used in Chapter 2.

Study	Accession number	Cell type	Target	Rep
Hainer 2018 [120]	GSE111121	ESC	Nanog	2
Hainer 2018 [120]	GSE111121	ESC	Sox2	2
Hainer 2018 [120]	GSE111121	ESC	Oct4	2

Table B.4: ChIP-seq datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	ChIP	Rep
Atlasi 2019 [199]	GSE92412	ESC 2i	Esrrb	2
Atlasi 2019 [199]	GSE92412	ESC serum	Esrrb	2
Atlasi 2019 [199]	GSE92412	ESC 2i	H3K27ac	1
Atlasi 2019 [199]	GSE92412	ESC 2i Nr5a2 KO	H3K27ac	1
Atlasi 2019 [199]	GSE92412	ESC 2i day3	H3K27ac	2
Atlasi 2019 [199]	GSE92412	ESC 2i day3 Esrrb KO	H3K27ac	2
Atlasi 2019 [199]	GSE92412	ESC 2i LT	H3K27ac	2
Atlasi 2019 [199]	GSE92412	ESC 2i LT Esrrb KO	H3K27ac	2
Engelen 2015 [386]	GSE58113	ESC	Dppa2	1
Kartikasari 2013 [199]	GSE44764	ESC	Tbx3	2
Steinhauser 2020	-	ESC serum	Nanog	2
Rafiee 2016 [204]	E-MTAB-3802	ESC 2i	Nanog	2
Cruz-Molina 2017 [231]	GSE89211	ESC 2i	H3K4me1	1
Cruz-Molina 2017 [231]	GSE89211	ESC 2i	H3K27ac	1
Cruz-Molina 2017 [231]	GSE89211	ESC 2i	H3K27me3	1
Cruz-Molina 2017 [231]	GSE89211	ESC 2i	p300	1
Cruz-Molina 2017 [231]	GSE89211	ESC serum	H3K4me1	1
Cruz-Molina 2017 [231]	GSE89211	ESC serum	H3K27ac	1
Cruz-Molina 2017 [231]	GSE89211	ESC serum	H3K27me3	1
Cruz-Molina 2017 [231]	GSE89211	ESC serum	p300	1
Galonska 2014 [293]	GSE53768	ESC	Utf1	1
Galonska 2015 [216]	GSE56312	ESC serum	Nanog	2
Galonska 2015 [216]	GSE56312	ESC 24h 2i	Nanog	2
Galonska 2015 [216]	GSE56312	ESC 3p 2i	Nanog	2
Galonska 2015 [216]	GSE56312	ESC serum	Sox2	2
Galonska 2015 [216]	GSE56312	ESC 24h 2i	Sox2	2
Galonska 2015 [216]	GSE56312	ESC 3p 2i	Sox2	2
Galonska 2015 [216]	GSE56312	ESC serum	Oct4	2
Galonska 2015 [216]	GSE56312	ESC 2i	Oct4	2
Novo 2018 [217]	GSE103053	EpiSC	Nanog	1
Sun 2018 [263]	GSE115340	ESC	Med1	2
Sun 2018 [263]	GSE115340	ESC	Med12	2
Denissov 2014 [387]	GSE52071	ESC	Ash2l	1
Li 2010 [388]	GSE19708	ESC	Jarid2	1
Marson 2008 [298]	GSE11724	ESC	Tcf7l1	2
Lim 2010 [299]	GSE37889	ESC	Zic3	2
Zhang 2013 [301]	GSE43565	ESC	Ctnnb1	2

Table B.5: ATAC-seq datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	Rep
Kearns 2015 [224]	GSE64059	ESC 2i	1
Atlasi 2019 [199]	GSE92412	ESC 2i 8h	3
Atlasi 2019 [199]	GSE92412	ESC 2i day1	3
Atlasi 2019 [199]	GSE92412	ESC 2i day3	3
Atlasi 2019 [199]	GSE92412	ESC serum 8h	3
Atlasi 2019 [199]	GSE92412	ESC serum day1	3
Atlasi 2019 [199]	GSE92412	ESC serum day3	3

Table B.6: GRO-seq datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	Rep
Tastemel 2017 [247]	GSE99747	ESC 2i	3
Tastemel 2017 [247]	GSE99747	ESC serum	2

Table B.7: Capture Hi-C datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	URL
Atlasi 2019 [199]	GSE72164	ESC 2i & serum	Link
Joshi 2015 [145]	GSE72164	ESC 2i & serum	Link
Sahlen 2015 [146]	GSE60495	ESC serum	Link

Table B.8: WGBS datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	Rep
Stadler 2011 [91]	GSE30202	ESC serum	1
Stadler 2011 [91]	GSE30202	NPC	1
Domcke 2015 [278]	GSE67867	ESC 2i to serum	1
Domcke 2015 [278]	GSE67867	ESC serum to 2i	1

Table B.9: RNA-seq datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	Rep
Joshi 2015 [145]	GSE72164	ESC 2i	2
Joshi 2015 [145]	GSE72164	ESC serum	2

Table B.10: Proteomics datasets used in Chapter 2 & 3.

Study	Accession number	Cell type	Rep	URL
Rafiee 2016 [204]	PXD003798	ESC 2i & serum	3	Link
van Mierlo 2019 [205]	PXD007154	ESC 2i& serum	3	Link

Table B.11: Other publicly available datasets used in Chapter 2 & 3.

Description	URL
Transcription factor database	http://humantfs.ccbr.utoronto.ca/
Cistrome database	http://cistrome.org/db/
LOLA region database	http://big.databio.org/regiondb
Eukaryotic promoter database	https://epd.epfl.ch//index.php

B.2 Chapter 4 Supplementary Tables

Table B.12: Published ATAC-seq datasets used in Chapter 4.

Study	Accession number	Cell type	Rep
Neijts 2016 [325]	GSE81203	embryo-day6.0	2
Neijts 2016 [325]	GSE81203	embryo-day7.2	2
Neijts 2016 [325]	GSE81203	embryo-day7.5-post	1

Table B.13: Published mRNA-seq datasets used in Chapter 4.

Study	Accession number	Cell type	Rep
Gouti2014 [310]	EMTAB2268	D1	3
Gouti2014 [310]	EMTAB2268	D2	3
Gouti2014 [310]	EMTAB2268	D3A	3
Gouti2014 [310]	EMTAB2268	D3NMP	3
Gouti2014 [310]	EMTAB2268	D4A	3
Gouti2014 [310]	EMTAB2268	D4H	3
Gouti2014 [310]	EMTAB2268	D4SC	3
Gouti2014 [310]	EMTAB2268	D5A	3
Gouti2014 [310]	EMTAB2268	D5H	3
Gouti2014 [310]	EMTAB2268	D5SC	3
Juan2016 [389]	GSE85717	ESC-serum	2

Table B.14: Published microarray datasets used in Chapter 4.

Study	Accession number	Cell type	Rep
Mazzoni 2013 [356]	GSE39422	motor neuron progenitors	3
Mazzoni 2013 [356]	GSE39422	motor neuron progenitors +iCdx2 +FGF	3

Table B.15: ChIP-seq datasets used in Chapter 4.

Study	Accession number	Cell type	ChIP	Rep
Aloia 2013 [390]	GSE53542	Neural progenitors	Zrf1	1
Amin 2017 [315]	GSE84899	NMP	Cdx2	2
Amin 2017 [315]	GSE84899	NMP	TBra	2
Bergsland 2011 [391]	GSE33059	Neuron	Sox11	3
Borromeo 2014 [392]	GSE55840	E12.5 mouse neural tube	Ascl1	1
Borromeo 2014 [392]	GSE55840	E12.5 mouse neural tube	Ptf1a	1
Borromeo 2014 [392]	GSE55840	E12.5 mouse neural tube	Rbpj	1
Jung 2010 [393]	GSE21812	motor neuron progenitors	Hoxc9	1
Koch 2017 [318]	GSE93524	NMP	Sox2	1
Koch 2017 [318]	GSE93524	NMP	T	1
Koch 2017 [318]	GSE93524	NMP	Tbx6	1
Koch 2017 [318]	GSE93524	NMP	betaCatenin	1
Mazzoni 2013 [333]	GSE31456	cortical motor neurons	Isl1-2	2
Mazzoni 2013 [333]	GSE31456	cortical motor neurons	Phox2a	2
Mazzoni 2013 [333]	GSE31456	motor neuron progenitors	Isl1-2	2
Mazzoni 2013 [333]	GSE31456	motor neuron progenitors	Lhx3	2
Mazzoni 2013 [356]	GSE39433	motor neuron progenitors	Cdx2-FGF	1
Mazzoni 2013 [356]	GSE39433	motor neuron progenitors	Cdx2-noFGF	1
Mcaninch 2014 [394]	GSE57186	Neural progenitors	Sox3	3
Mistri 2015 [395]	GSE69859	neural stem cells	Brn1	1
Mistri 2015 [395]	GSE69859	neural stem cells	Brn2	1
Mistri 2015 [395]	GSE69859	neural stem cells	Pou3f1	1
Mistri 2015 [395]	GSE69859	neural stem cells	Sox2	1
Nishi 2012 [**Nishi2012**]	GSE42132	neural progenitors (from embryoid bodies)	Foxa2	1
Nishi 2012 [**Nishi2012**]	GSE42132	neural progenitors (from embryoid bodies)	H3K27ac-RA	1
Nishi 2012 [**Nishi2012**]	GSE42132	neural progenitors (from embryoid bodies)	H3K27ac-SAG	1
Nishi 2015 [396]	GSE61673	Neural progenitors	Gli3	1
Nishi 2015 [396]	GSE61673	Neural progenitors	Nkx2.2	2
Nishi 2015 [396]	GSE61673	Neural progenitors	Nkx6.1	2
Nishi 2015 [396]	GSE61673	Neural progenitors	Olig2	2
Metzis 2018 [305]	E-MTAB-6348	Day 5 hindbrain neural progenitor	Sox2	2
Metzis 2018 [305]	E-MTAB-6348	Day 5 spinal cord progenitor	Sox2	2

Table B.16: Other publicly available datasets used in Chapter 4.

Description	URL
Vista enhancer set	https://www.encodeproject.org
4DGenome interaction data	https://4dgenome.research.chop.edu
LOLA region database	http://big.databio.org/regiondb
ENCODE DNAse hypersensitive sites	https://www.encodeproject.org