

Regulation of cell-type specific enhancers in mouse pluripotent stem cells

+

Emma Bell

+

Thesis submitted towards fulfilment of the degree of
Doctor of Philosophy

+

Department of Surgery and Cancer
Faculty of Medicine
Imperial College London

Declaration of originality

Unless otherwise stated, the author conducted all work described in this thesis. The following people made significant contributions:

Name	Contribution
Prof Ian Chambers	<ul style="list-style-type: none"> • Provided the ESC^{Nanog^{-/-}} and ESC^{Nanog^{+/+}} lines.
Prof Austin Cooney	<ul style="list-style-type: none"> • Provided the ESC^{Ncoa3^{-/-}} and ESC^{Ncoa3^{+/+}} lines.
Dr Edward Curry	<ul style="list-style-type: none"> • Devised and wrote the <i>plotCoverage</i> R function to calculate the mean sequencing read coverage across a set of genomic regions.
Dr Alice Jouneau	<ul style="list-style-type: none"> • Provided the EpiSC bisulphite-seq data. • Provided the ESC and EpiSC RNA-seq data. • Performed the ESC-to-EpiSC conversion and accompanying RT-qPCR.
Dr Luc Jouneau	<ul style="list-style-type: none"> • Processed and analysed the ESC and EpiSC bisulphite-seq data.
King Hang Mau	<ul style="list-style-type: none"> • Cultured the ESC cell lines bearing homozygous deletions of <i>Esrrb</i>, <i>Ncoa3</i>, and <i>Nanog</i> and wild-type controls. • Performed RT-qPCR of these same cells.
Dr Austin Smith	<ul style="list-style-type: none"> • Provided the ESC^{Esrrb^{-/-}} and ESC^{Esrrb^{+/+}} lines.
Dr Rute Tomaz	<ul style="list-style-type: none"> • Generated ChIP libraries from ESC^{Esrrb^{-/-}} and ESC^{Esrrb^{+/+}}. • Performed the rescue experiments of ESC^{Esrrb^{-/-}} and ESC^{Esrrb^{+/+}}.
Dr Wout Megchelenbrink	<ul style="list-style-type: none"> • Analysed the HiC-seq data validating Super Enhancer-promoter interactions in ESCs.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Abstract

The hundreds of different cell types in the human body arise from a single cell: the zygote. This ability is called totipotency. As a cell becomes more specialised its potency narrows, reducing the number and diversity of its daughter cells. We can derive two pluripotent stem cell populations from mouse pre- and post-implantation embryos: Embryonic Stem Cells (ESCs) and Epiblast Stem Cells (EpiSCs). These cell types respectively represent the naive and primed states of the pluripotency spectrum. A delicately balanced transcriptional network underlies ESCs and EpiSCs, capable of both preserving pluripotency, yet prepared for differentiation. Enhancers act as hubs of transcription factors to focus and promote gene expression. In 2013 Prof Rick Young's group proposed an ostensibly novel enhancer class: Super Enhancers (SE). In contrast to mere Typical Enhancers (TE), SE purportedly recruit transcription factors at extremely high density to drive strong expression of cell-specific genes.

This thesis examines the regulation and activity of Super Enhancers defined in ESCs during pluripotency maturation. To achieve this, we used an integrated, data-driven approach, analysing sequencing data pertaining to transcription factor occupancy, chromatin accessibility, epigenetic signatures, mRNA expression, and DNA-DNA interactions, to assay SE activity both in vitro and in vivo. We found that the constituent enhancers within SEs divide into two sub-populations based on whether they continue to function as primed pluripotency establishes. We termed the constituent enhancer units Persistently Unmethylated (PU) and Differentially Methylated (DM) Super Enhancelets, reflecting their CpG methylation profiles in naive and primed pluripotent cells. PU and DM enhancers harbour distinct sequence compositions, transcription factor occupancy, and epigenetic profiles. Furthermore DM exhibit a specific chromatin function for the orphan nuclear receptor ESRRB and its co-regulators. We conclude that ESC enhancers represent a heterogeneous population of genetic regulatory elements, controlled by distinct transcriptional mechanisms.

Acknowledgements

In *Mostly Harmless* by Douglas Adams, the protagonist Arthur Dent discovers that in the absence of human civilization he has little to offer the universe. In fact, "Left to his own devices he couldn't build a toaster. He could just about make a sandwich and that was it."⁽¹⁾ Similarly, without the guidance, support and love of a great many people, this thesis wouldn't have been much more than an unappetising 251-layer sandwich.

Firstly, I thank my primary supervisor Dr Veronique Azuara for her steadfast scientific guidance and teaching me so much over the past four years. I thank Dr James Flanagan for understanding and getting me through this. I thank Dr Ed Curry for all the tea, teaching me to code, and believing in me. I'd also like to thank Dr Curry's ridiculous cats Watson and Crick. They were the most congenial of colleagues, albeit a little food-driven at times. I thank the members – current and former – of the Epigenetics & Development lab and Epigenetics group, as well as my friends in the IRDB for their support and good humour, particularly Dr Rute Tomaz, Donja Karimlou, Dr Shuchen Zhang, and Dr Zainab 'Zizi' Al-Shareef.

As well as to my supervisors and peers, I owe an enormous debt of gratitude to my friends and loved ones. One in two PhD students develop depression and anxiety during the course of their research,⁽²⁾ as opposed to one in four people in the general population.⁽³⁾ I am one of those PhD students. I feel incredibly fortunate to have such a loving, extensive and unwavering support network. For all the ways you helped me, whether or not you realised you were doing so, I thank Emma Norris, Lucy Abel, Jonathan Sutton, Remi Friedman, Dai Brown, Bob Churchill, Dr Clare Huxley, and Pippa Mitchell.

The empathy and advice shared by those who've also undergone this experience – especially those who were writing up at the same time as me – made this process infinitely easier. I therefore thank Soon-To-Be Dr Amy Barstow, Soon-To-Be Dr Marie Holt, Soon-To-Be Dr David Corcoran, Dr Hanna Gillespie-Gallery, Dr Simon Barnes-Sadler, and Dr Mike Paynter.

For showing me what women can build together and letting me be a part of that, I thank my beloved girl gang Lips Choir. You've given me a great deal more than I can articulate here, but I know that you understand through the power of TheWeb™. During write-up I stuck my Pure Shores Holiday Camp Award on the wall by my desk. Thank you Amy Harris-Sandström for nominating me for the 'Gene-ius Award'; you gave me courage when I needed it.

On the subject of badass humans, I thank the glorious Eliza Goroya and the Wake the Witches Assembly for providing judgement-free emotional support, a safe place to learn and create, and helping me grow as a queer woman of colour ("Our backs tell stories no book has the spine to carry" – Rupī Kaur).

I thank Rory 'Quincy' Fenton for that one time he gave me his spare ticket to that recording of *Have I Got News For You* in January 2015. (See, I told you I would.)

Lastly, I thank Howard Shore for providing the soundtrack to every revision session and piece of course work that I have undertaken since I was 14 years old.

I will never be able to repay the debt that I owe to you all, so instead I dedicate to you this thesis. Fingers crossed at least some of it remains relevant in another four years.

Table of contents

Declaration of originality	i
Copyright Declaration	ii
Abstract	iii
Acknowledgements	iv
Table of contents	v
Figures	x
Tables	xiii
Abbreviations and acronyms	xv
1 Introduction	1-19
1.1 Preface	1-19
1.2 Early murine embryogenesis	1-19
1.1.1 Pre-implantation embryogenesis	1-19
1.1.2 The first steps of post-implantation embryogenesis	1-21
1.3 Modelling pluripotency with embryo-derived stem cells	1-21
1.1.3 Isolation and maintenance of stem cells from the early embryo	1-21
1.1.4 Pluripotent stem cell heterogeneity and plasticity	1-23
1.4 Transcriptional regulation of pluripotency	1-24
1.1.5 The OCT4-SOX2-NANOG triad	1-25
1.1.6 Naïve and primed pluripotency transcription factors	1-25
1.5 Epigenetic regulation of pluripotency	1-26
1.1.7 Chromatin organisation	1-26
1.1.8 The molecular basis of epigenetics	1-27
1.1.8.1 DNA methylation	1-27
1.1.8.2 Post-translational modifications of histone proteins	1-29
1.1.9 Epigenetic regulation of pluripotency	1-30
1.6 Enhancers	1-32
1.1.10 The early history of enhancer research	1-32
1.1.11 The capacity and mechanisms of enhancer function	1-34
1.1.12 The topological behaviour of enhancers	1-36
1.1.13 Super Enhancers and Typical Enhancers	1-37
1.7 Hypothesis and aims	1-38
2 Materials & Methods	2-40
2.1 Materials	2-40
2.1.1 Buffers, solutions and gels	2-40
2.1.2 Chemicals, solvents and reagents	2-43

2.1.3	Antibiotics	2-44
2.1.4	Antibodies	2-44
2.1.4.1	Primary antibodies	2-44
2.1.4.2	Secondary antibodies	2-44
2.1.4.3	Standard IgG antibodies	2-45
2.1.5	Cell lines	2-45
2.1.6	Enzymes	2-45
2.1.7	Kits	2-45
2.1.8	Inhibitors	2-46
2.1.8.1	Protease inhibitors	2-46
2.1.8.2	Phosphatase inhibitors	2-46
2.1.8.3	Signalling pathway inhibitors	2-46
2.1.9	Plasmids	2-46
2.1.10	Tissue culture components	2-46
2.1.10.1	Media recipes	2-46
2.1.10.1.1	ESC media	2-46
2.1.10.1.2	EpiSC media	2-47
2.1.10.1.3	Other	2-47
2.1.10.2	Media supplements	2-48
2.1.10.3	Other cell culture reagents	2-48
2.1.10.4	Culture vessel coatings	2-48
2.1.11	Resources and websites	2-48
2.1.11.1	Computer software	2-48
2.1.11.2	Websites	2-49
2.2	Methods	2-50
2.2.1	Wet lab methods	2-50
2.2.1.1	Cell culture	2-50
2.2.1.1.1	Embryonic stem cell culture	2-50
2.2.1.1.2	Epiblast stem cell culture	2-51
2.2.1.1.3	Cryopreservation	2-51
2.2.1.1.4	Producing LIF	2-51
2.2.1.2	Embryonic stem cell phenotype assays	2-52
2.2.1.2.1	Self-renewal assay	2-52
2.2.1.2.2	Colony formation efficiency assay	2-52
2.2.1.3	In vitro conversion from ESCs to EpiSCs	2-53
2.2.1.4	RT-qPCR	2-53
2.2.1.4.1	RNA extraction	2-53
2.2.1.4.2	cDNA synthesis	2-54

2.2.1.4.3	RT-qPCR	2-54
2.2.1.4.4	Primer design and validation	2-55
2.2.1.5	ChIP-seq	2-55
2.2.1.5.1	ChIP	2-55
2.2.1.5.2	Generation of libraries for sequencing.....	2-56
2.2.2	Computational methods	2-57
2.2.2.1	Sequence composition analysis	2-57
2.2.2.1.1	GC content	2-57
2.2.2.1.2	CpG content.....	2-57
2.2.2.2	Data collection.....	2-57
2.2.2.3	Microarray data	2-58
2.2.2.3.1	Microarray processing.....	2-58
2.2.2.3.2	Meta-analysis of microarray data	2-58
2.2.2.3.3	Gene enrichment analysis of microarray data	2-59
2.2.2.4	Sequencing data	2-60
2.2.2.4.1	Sequencing data processing.....	2-60
2.2.2.4.1.1	Sequencing read processing	2-60
2.2.2.4.1.2	Bisulphite-seq	2-60
2.2.2.4.1.3	ChIP-seq	2-60
2.2.2.4.1.4	DIP-seq	2-61
2.2.2.4.1.5	RNA-seq.....	2-61
2.2.2.4.2	Sequencing data analysis	2-62
2.2.2.4.2.1	Gene annotation and motif enrichment	2-62
2.2.2.4.2.2	Read coverage comparison	2-62
2.2.2.5	General analyses	2-63
2.2.2.5.1	Gene ontology	2-63
2.2.2.5.2	Gene enrichment analysis.....	2-63
2.2.2.5.3	Clustering analysis.....	2-63
3	Results I.....	3-64
3.1	Introduction	3-64
3.2	Results.....	3-67
3.2.1	ESC Super Enhancers and Typical Enhancers appear CpG methylated in EpiSCs	3-67
3.2.2	Dormant Super Enhancers appear hypermethylated in ESCs.....	3-72
3.2.3	Transcription factor binding at ESC Super Enhancers clusters at hypomethylated constituent enhancers.....	3-74
3.2.4	ESC Super Enhancers physically interact with their proposed associated genes in ESCs	3-79

3.2.5	ESC Super Enhancer associated gene promoters acquire CpG methylation during pluripotency maturation	3-79
3.2.6	Post-implantation epiblast cells <i>in vitro</i> and <i>in vivo</i> downregulate transcription of ESC Super Enhancer associated genes	3-82
3.2.7	Genes associated with ESC Super Enhancer compared to ESC Typical Enhancers appear more sensitive to the loss of naïve pluripotency factors	3-88
3.3	Conclusions	3-89
4	Results II.....	4-92
4.1	Introduction	4-92
4.2	Results.....	4-93
4.2.1	ESC Super Enhancers acquire CpG methylation in distinct patterns in EpiSCs	4-93
4.2.2	ESC Super Enhancelets selectively acquire CpG methylation in epiblast <i>in vivo</i>	4-99
4.2.3	ESC Super Enhancelets display distinct methylation patterns in non-pluripotent cell types.....	4-100
4.2.4	Persistently Unmethylated and Differentially Methylated ESC Super Enhancelets display distinct epigenetic profiles in ESCs	4-103
4.2.5	Persistently Unmethylated and Differentially Methylated enhancer units display different histone modification profiles during pluripotency maturation	4-108
4.2.6	Differentially Methylated relative to Persistently Unmethylated enhancer units lose transcription factor occupancy during pluripotency maturation.....	4-112
4.2.7	The selective retention of unmethylated Super Enhancelets defines two distinct classes of ESC Super Enhancers with differential activity in EpiSCs and epiblast <i>in vivo</i>	4-114
4.2.8	Class I and II super-enhancer associated genes show differential expression patterns upon ESC-to-EpiSC conversion <i>in vitro</i>	4-120
4.3	Conclusion	4-124
5	Results III.....	5-126
5.1	Introduction	5-126
5.2	Results.....	5-127
5.2.1	shRNA screen identifies Silenced ESC Super Enhancer associated genes most sensitive to the loss of naïve pluripotency factors.....	5-127
5.2.2	ERRE and KLF motifs appear enriched at DM Super Enhancelets	5-137
5.2.3	ESRRB and NCOA3 selectively bind to DM ESC Super Enhancelets at high density relative to other naïve and primed pluripotency factors	5-141
5.2.4	Meta-analysis of microarray data evidences that Silenced ESC Super Enhancer associated genes are sensitive to the loss of ESRRB	5-144
5.2.5	ESRRB requires its essential co-activator NCOA3 to mediate its regulation of Silenced Super Enhancers associated genes.	5-153
5.2.6	ESRRB loss selectively depletes MED1 at DM Super Enhancelets	5-161
5.3	Conclusion	5-166

6	Discussion	6-169
6.1	The signal: ESC Super Enhancers and Typical Enhancers during pluripotency maturation	6-169
6.1.1	Do ESC Super Enhancers and Typical Enhancers retain the same open epigenetic signature in EpiSCs?	6-169
6.2	The subdivision: the two classes of Super Enhancer in ESCs	6-174
6.2.1	Some ESC Super Enhancer constituent enhancer units maintain an open chromatin state during pluripotency maturation	6-174
6.2.2	During pluripotency maturation, PU-containing ESC Super Enhancers maintain protein binding and transcription of their associated genes	6-175
6.3	The fulcrum and the lever: H3K4me3 balances TET/DNMT at PU and DM.....	6-177
6.4	The linchpin: DM activity relies on ESRRB protein and co-regulators.....	6-180
6.4.1	Does the stability of DM rely on the presence of ESRRB-NCOA3?.....	6-182
6.4.2	Does ESRRB form part of a DM-specific regulatory network?.....	6-183
6.5	The beacon: enhancers in murine pluripotent cell types.....	6-185
	Bibliography	189
A	Appendices	218
1.1	Tables.....	218
1.2	Code.....	239
1.2.1	Sequencing data processing.....	239
1.2.1.1	Single-end sequencing reads	239
1.2.1.2	Paired end sequencing reads	241
1.2.1.3	Generating coverage plots from bam files	243
1.2.2	Microarray processing.....	245
1.2.2.1	Affymetrix arrays	245
1.2.2.2	Agilent arrays	246
1.2.2.3	Illumina arrays	247
1.2.2.4	Compiling a table of t-statistics for each annotated gene in the mouse genome	248
1.2.2.5	Meta-analyses	249
1.2.2.5.1	Fisher's method	249
1.2.2.5.2	Rank products.....	250
1.2.2.6	Gene set enrichment	251

Figures

Chapter 1: Introduction

Figure 1.1 Pre-implantation murine embryogenesis.	1-20
Figure 1.2 The first steps of post-implantation embryonic development	1-21
Figure 1.3 Deriving pluri- and multipotent stem cells from the developing murine embryo.....	1-22
Figure 1.4 The heterogeneity of in vitro pluripotency.	1-24
Figure 1.5 Epigenetic regulation of pluripotency.	1-32
Figure 1.6 Enhancer research over time.	1-34
Figure 1.7 ESC Super Enhancers act as turbo-charged cell-specific enhancers.	1-39

Chapter 2: Materials and methods

Figure 2.1 An overview of the sequencing read processing pipeline.	2-60
---	------

Chapter 3: Results I

Figure 3.1 ESC Super Enhancers and Typical Enhancers acquire CpG methylation as pluripotency matures.....	3-69
Figure 3.2 ESC but not ProB Super Enhancers appear selectively CpG hypomethylated in ESCs.	3-74
Figure 3.3 CpG hypomethylation delineates the constituent enhancers within ESC Super Enhancers.....	3-76
Figure 3.4 ESC Super Enhancers physically interact with their proposed associated genes. ...	3-80
Figure 3.5 ESC Super Enhancer associated gene promoters acquire CpG methylation during pluripotency maturation.....	3-81
Figure 3.6 Epiblast cells, in vitro and in vivo, downregulate transcription of ESC Super Enhancer associated genes.	3-85
Figure 3.7 ESC Super Enhancer associated genes selectively deregulate during pluripotency maturation.....	3-89

Chapter 4: Results II

Figure 4.1 ESC Super Enhancers acquire CpG methylation in distinct patterns in EpiSCs.....	4-98
Figure 4.2 ESC Super Enhancers acquire CpG methylation in distinct patterns in epiblast in vivo.	4-101
Figure 4.3 ESC Super Enhancelets display distinct CpG methylation profiles in non-pluripotent cell types.....	4-105
Figure 4.4 ESC Super Enhancelets display distinct CpG methylation profiles in ESCs. .	4-107
Figure 4.5 ESC Super Enhancelets harbour different histone modification profiles in pluripotent stem cells.	4-111
Figure 4.6 PU compared to DM ESC Super Enhancelets bind higher densities of pluripotency transcription factors.....	4-113
Figure 4.7 Only genes associated with Class II ESC Super Enhancers tend to reduce transcription as pluripotency matures.....	4-119
Figure 4.8 Class II ESC Super Enhancer associated genes gradually reduce transcription during ESC-to-EpiSC conversion, whereas Class I maintain expression.....	4-124

Chapter 5: Results III

Figure 5.1 Identifying candidate transcription factors that selectively regulate PU and DM ESC Super Enhancelet activity using a publicly available shRNA screen.	5-136
Figure 5.2 DNA binding motifs selectively enriched at PU and DM Super Enhancelets..	5-138
Figure 5.3 Protein binding density at PU and DM Super Enhancelets of identified candidate regulators using publicly available ChIP-seq data.	5-143
Figure 5.4 Candidate differential regulators of ESC Super Enhancelets.....	5-144
Figure 5.5 Phenotypic and functional characterisation of ESCs bearing homozygous deletions of Esrrb and Ncoa3.	5-147
Figure 5.6 Meta-analysis of microarray experiments performed on embryonic stem cells depleted for Esrrb and Nanog.	5-153

Figure 5.7 Comparative RT-qPCR analysis of Maintained and Silenced Super Enhancer associated gene expression in embryonic stem cells constitutively depleted for Esrrb, Ncoa3 or Nanog.	5-157
Figure 5.8 The generation of ESC ^{Esrrb-/-} rescue cells.....	5-158
Figure 5.9 ESRRB requires a functional AF2 domain to positively regulate the expression of Silenced Super Enhancer associated genes in ESCs.	5-161
Figure 5.10 Assessing the impact of ESRRB loss on H3K27a deposition and MED1 binding at DM and PU Super Enhancelets in ESC.....	5-164

Chapter 6: Discussion

Figure 6.1 Epigenetic remodelling of enhancers rather than promoters allows multiple enhancers to drive the same gene under different transcriptional programmes.....	6-172
Figure 6.2 The four classes of ESC enhancer.....	6-176
Figure 6.3 The H3K4me3-TET-DNMT axis.	6-179
Figure 6.4 The differing balance between ESRRB-NCOA3 and core transcriptional regulators at PU and DM..	6-181
Figure 6.5 The formal (PubMed and bioRxiv) and informal (Google searches and tweets) discussion of Super Enhancers.....	6-187

Tables

Chapter 3: Results I

Table 3.1 ESC Super Enhancer gene promoters that are CpG hypermethylated in ESCs and EpiSCs.....	3-82
Table 3.2 Super Enhancer associated genes by their differential expression in EpiSCs and ex vivo epiblast compared to ESCs.	3-86
Table 3.3 Gene Ontology terms enriched for of ESC Super Enhancer associated genes.....	3-87

Chapter 4: Results II

Table 4.1 Gene Ontology terms enriched in Class I and II ESC Super Enhancer associated gene sets.	4-117
--	-------

Chapter 5: Results III

Table 5.1 Super Enhancer associated genes for which the microarray platform used by Nishyama et al. had no target probes.....	5-129
Table 5.2 The expression of ESC Super Enhancer associated genes upon shRNA inhibition of pluripotency transcription factors.....	5-133
Table 5.3 HOMER motif analysis of ESC Super Enhancers.....	5-139
Table 5.4 Genes that appear differentially expressed in the absence of Esrrb or Nanog according to Rank Products meta-analysis of microarray data.	5-149
Table 5.5 Rank Products designates Esrrb and Nanog as statistically significantly downregulated in the Esrrb and Nanog depleted cell lines compared to their normal ESC controls.	5-150
Table 5.6 The expression of Maintained and Silenced Super Enhancer associated genes in ESCs bearing homozygous deletions of Esrrb, Nco3 or Nanog compared to their wild-type controls.	5-157

Table 5.7 The expression of Maintained and Silenced Super Enhancer associated genes in ESCs bearing a homozygous deletion of Esrrb, rescued with either wild-type Esrrb or mutant Esrrb, compared to their wild-type controls.	5-162
---	-------

Table 5.8 Candidate ESC Super Enhancelet differential regulators.....	5-167
---	-------

Chapter 6: Discussion

Table 6.1 OCT4-independent ESRRB interaction partners.	6-184
---	-------

Appendices

Table A.1 RT-qPCR primer sequences.	218
--	-----

Table A.2 GEO indexed sequencing and microarray data sets used in this study.	218
--	-----

Table A.3 The ChIP-seq data visualised in figure 3.2A.....	221
--	-----

Table A.4 ESC Super Enhancers and their associated genes, by Super Enhancelet complementation and class.....	222
--	-----

Table A.5 Gene Ontology terms enriched in the Class I (Maintained) ESC Super Enhancer associated gene set.	226
---	-----

Table A.6 Gene Ontology terms enriched in the Class II (Silenced) ESC Super Enhancer associated gene set.	232
--	-----

Table A.7 Mean-rank gene set test of Class I and II ESC Super Enhancer associated genes.	235
---	-----

Table A.8 The differential expression of shRNA targets in Nishyama et al.'s ESC shRNA screen.....	237
---	-----

Abbreviations and acronyms

2i - 2 Inhibitors,	1-6	EGTA - Ethylene Glycol Tetraacetic Acid,	2-26
5-caC - 5-carboxylcytosine,	1-10	EPC - Ectoplacental Cone,	1-3
5-fC - 5-formylcytosine,	1-10	EpiLC - Epiblast-Like Cell,	4-78
5-hmC - 5-hydroxymethylcytosine,	1-10	EpiSC - Epiblast Stem Cell,	1-4
5-mC - 5-methylcytosine,	1-9	EPU - Enhancer-Promoter Unit,	6-154
APS - Ammonium Persulfate,	2-26	eRNA - Enhancer RNA,	6-154
BE - Boundary Element,	1-18	ESC - Embryonic Stem Cell,	1-4
BMP - Bone Morphogenetic Protein, ...	1-4	ExE - Extra-embryonic ectoderm,	4-85
BSA - Bovine Serum Albumin,	2-26	ExE - Extra-Embryonic Ectoderm,	1-3
CGI - CpG Island,	1-9	FAIRE-seq - Formaldehyde-Assisted Isolation of Regulatory Elements Sequencing,	4-97
CPDB - Consensus Path Database, ..	2-46	FC - Fold-Change,	3-66
CpG - Cytosine-Guanine dinucleotide, ..	1-9	FDR - False Discovery Rate,	2-42
DIP-seq - DNA Immunoprecipitation Sequencing,	2-44	fMeDIP - Formylcytosine Methylated DNA Immunoprecipitation,	2-44
DM - Differentially Methylated,	4-76	GEO - Gene Expression Omnibus, ...	2-40
DMSO - Dimethyl Sulfoxide,	2-26	gRNA - guide RNA,	1-10
DNMT - DNA Methyltransferase,	1-9	HAT - Histone Acetyltransferase,	1-11
dNTP - Deoxynucleotide Triphosphate,	2-26	HDAC - Histone Deacetylase,	1-12
DOC - Deoxycholic acid,	2-26	HKMT - Histone Lysine Methyltransferase,	1-12
DTT - DL-Dithiothreitol,	2-26	hMeDIP - Hydroxymethylated DNA Immunoprecipitation,	2-44
E - embryonic day,	1-1		
EDTA - Ethylene Diamine Tetraacetic Acid,	2-26		

HMM - Hidden Markov Model,	2-43	PM - Persistently Methylated,	4-77
HSC - Haematopoietic Stem Cell,	1-1	PrE - Primitive Endoderm,	1-2
ICM - Inner Cell Mass,	1-2	PU - Persistently Unmethylated,	4-76
iPSCs - induced Pluripotent Stem Cells, ..	3-47	RMS - Relative Methylation Score, ...	2-44
LCR - Locus Control Region,	6-169	RNAi - RNA interference,	2-41
LE - Luminal Epithelium,	1-3	RP - Rank Products,	5-132
LEF - Loop Extrusion Factor,	1-18	RPKM - Reads Per Kilobase of transcript	
LIF - Leukaemia Inhibitory Factor,	1-4	per Million mapped reads,	4-98
MBD - Methyl Binding Domain,	1-11	RRBS - Reduced Representation	
mCpG - Methylated CpG dinucleotide, 1-9		Bisulphite Sequencing,	4-82
MeCP2 - Methyl-CpG-Binding Protein	2, 1-12	RS - Rank Sum,	5-132
MeDIP - Methylated DNA		SAM - S-Adenosyl Methionine,	1-9
Immunoprecipitation,	2-44	SDS - Sodium Dodecyl Sulphate,	2-27
MEF - Mouse Embryonic Fibroblast, ..	2-33	SRA - Sequence Read Archive,	2-43
MGI - Mouse Genome Index,	2-41	SV40 - Simian Vacuolating Virus 40, .	1-15
MTL - Multi Transcription Factor Loci,	6-169	TAD - Topological Associated Domains, ..	1-18
NP-40 - Nonidet P-40,	2-26	TCA - Trichloroacetic Acid,	2-27
OSN - OCT4-SOX2-NANOG,	1-7	TE - Trophectoderm,	1-2
PBS - Phosphate-buffered Saline,	2-27	TRD - Transcriptional Repressor Domain,	1-12
PCA - Principal Component Analysis, ...	2-46	TS - Trophoblast stem cell,	1-4
PE - Parietal Endoderm,	1-3	VE - Visceral Endoderm,	1-3
PFA - Paraformaldehyde,	2-26	WGBS - Whole Genome Bisulphite	
PIC - Promoter Initiation Complex,	1-18	Sequencing,	4-82
		XEN - Extra-embryonic endoderm,	1-4

One feels even in the midst of the traffic, or waking at night, Clarissa was positive, a particular hush, or solemnity; an indescribable pause; a suspense before Big Ben strikes. There! Out it boomed. First a warning, musical; then the hour, irrevocable. The leaden circles dissolved in the air. Such fools we are, she thought, crossing Victoria Street. For Heaven only knows why one loves it so, how one sees it so, making it up, building it round one, tumbling it, creating it every moment afresh; but the veriest frumps, the most dejected of miseries sitting on doorsteps (drink their downfall) do the same; can't be dealt with, she felt positive, by Acts of Parliament for that very reason: they love life. In people's eyes, in the swing, tramp, and trudge; in the bellow and the uproar; the carriages, motor cars, omnibuses, vans, sandwich men shuffling and swinging; brass bands; barrel organs; in the triumph and the jingle and the strange high singing of some aeroplane overhead was what she loved; life; London; this moment in June.

Virginia Woolf, Mrs. Dalloway

1 INTRODUCTION

1.1 Preface

The hundreds of different cell types in a human body all arise from a single cell: the zygote. This ability is called totipotency from the Latin 'tōtus', meaning whole or entire and 'potential' referring to capacity or possibility.⁽⁴⁾ As a cell becomes more specialised its potency narrows, reducing the number and diversity of the different types of daughter cell that it can give rise to. The transition from totipotent to pluripotent represents the first step in this narrowing of potency. Pluripotent cells may produce any cell type of the embryo, but unlike totipotent cells they cannot form extra-embryonic tissues such as the placenta. Multipotent or progenitor cells can only produce daughter cells of a specific lineage. For example, Haematopoietic Stem Cells (HSCs) only generate the different types of blood cell.

Despite sharing an identical genome, cells exhibit diverse functions and morphology as a result of specific gene expression profiles. The work here described uses *in vitro* cultures of pluripotent embryo-derived stem cells to model the behaviour of the genetic regulatory elements called enhancers, which promote the expression of both ubiquitous and cell-specific gene expression profiles. In this chapter, I first introduce mouse embryonic development, the concept of pluripotency, and detail its transcriptional regulation. I then give a historical and contemporary overview of research into enhancers, focusing on their role in pluripotent embryo-derived stem cells.

1.2 Early murine embryogenesis

1.1.1 Pre-implantation embryogenesis

In the mouse embryo the transition from totipotency to pluripotency occurs at approximately embryonic day (E) 3 with the formation of the first differentiated cell types. The first cell division takes place approximately 16-20 hours after fertilisation generating two identical blastomeres. This process continues until approximately E2.5 when the 8-cell embryo undergoes compaction, producing a dense embryo where each cell melts into its neighbours. At E3.0 the

embryo enters the morula stage, where blastomeres begin to take on different properties depending on whether located in the centre or at the periphery. After the fifth cell division, some blastomeres begin to exocytose vesicles into the embryo, which fuse together creating an increasingly large fluid filled space called the blastocoel. The external cells assume an epithelial morphology and extra-embryonic identity termed the trophectoderm (TE). The TE goes on to form the placenta. Meanwhile, the internal cells form a clump on one side of the embryo attached to the basal surface of the TE, termed the Inner Cell Mass (ICM). Cells of the ICM harbour pluripotency. By E4.5 the ICM cells exposed to the blastocoel become primitive endoderm (PrE), a second extra-embryonic lineage, which ultimately becomes the endodermal surface of the extra-embryonic yolk sac. Both the TE and PrE exhibit multipotency, as they can only give rise to cells of the placenta and yolk sac. Concurrently, the ICM becomes the epiblast, a pluripotent tissue that ultimately gives rise to the embryo proper (figure 1.1).

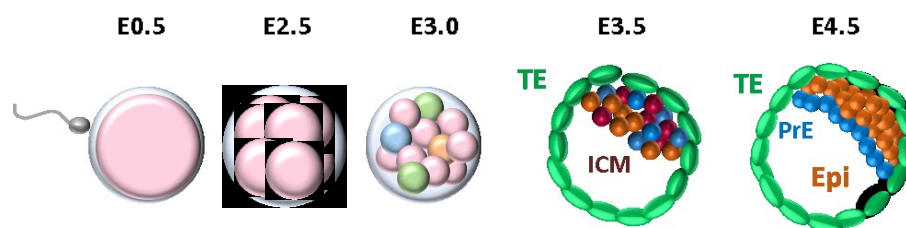


Figure 1.1 Pre-implantation murine embryogenesis.

16-20 hours after fertilisation the zygote undergoes its first cell division, producing two identical blastomeres. At embryonic day (E) 2.5, the embryo comprises eight compacted blastomeres. At E3.0 the embryo becomes a morula and the blastomeres begin to show different properties depending on their location within the embryo. By E3.5 the cells at the periphery of the embryo become to extra-embryonic trophectoderm (TE) and the cell at the centre of the embryo become the pluripotent Inner Cell Mass (ICM). At E4.5 the cells of the ICM exposed to the blastocoel become primitive endoderm (PrE), whereas the remainder mature to the pluripotent epiblast (Epi). The epiblast eventually gives rise to the embryo proper. Figure adapted from Tomaz, R.(228)

1.1.2 The first steps of post-implantation embryogenesis

Implantation refers to the event of the embryo fusing to the uterine wall. Up until this point the embryo has been travelling through the fallopian tube to emerge into the uterus. The TE fuses with the Luminal Epithelium (LE) of the uterine wall, forming the Ectoplacental Cone (EPC). Once the embryo achieves implantation, the epiblast undergoes drastic remodelling, shaping itself into a rosette-like cylinder containing a pro-amniotic cavity. The extra-embryonic tissues further differentiate. The TE becomes the Extra-Embryonic Ectoderm (ExE). The PrE forms the Parietal Endoderm (PE) and the Visceral Endoderm (VE), which enclose the epiblast (figure 1.2).

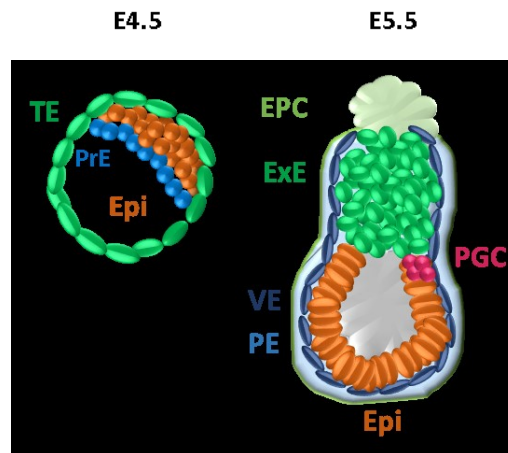


Figure 1.2 The first steps of post-implantation embryonic development

The TE fuses with the Luminal Epithelium (LE) of the uterine wall, becoming to Ectoplacental Cone (EPC). The embryo drastically remodels, forming a rosette-like cylinder. The TE matures to the Extra-Embryonic Ectoderm (ExE). The PrE forms the Parietal (PE) and Visceral Endoderms (VE),enveloping the embryo. Figure adapted from Tomaz, R.(228)

1.3 Modelling pluripotency with embryo-derived stem cells

1.1.3 Isolation and maintenance of stem cells from the early embryo

Pluripotent progenitors lack the ability to generate extra-embryonic tissues, but can differentiate into all three germ layers (ectoderm, mesoderm and endoderm) and primordial

germ cells, which eventually become the organism's gametes. Within the developing mouse embryo there exist two sources of pluripotent stem cells; embryonic stem cells (ESCs) arise from the pre-implantation E3.5-4.5 ICM and epiblast stem cells (EpiSCs) from the post-implantation E5.5-E8 epiblast. Despite the wide array of time-points from which one can derive EpiSCs, *in vitro* they appear highly similar.⁽⁵⁾ In addition to pluripotent stem cells, one can derive extra-embryonic stem cell lines from the same time-points. Extra-embryonic endoderm (XEN) stem cells derive from the PrE. Trophoblast stem cells (TS) derive from the TE. Both XEN and TS possess multipotency (figure 1.3).

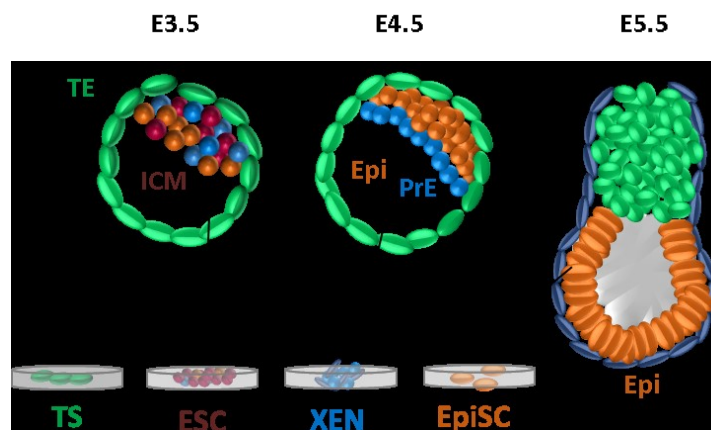


Figure 1.3 Deriving pluri- and multipotent stem cells from the developing murine embryo.

One can derive two types of pluripotent stem cell from the developing murine embryo: embryonic stem cells (ESCs) from the ICM and epiblast stem cells (EpiSCs) from the epiblast. Additionally, one can derive multipotent, extra-embryonic stem cells: trophoblast stem cells (TS) from the TE and extra-embryonic endoderm stem cells (XEN) from the PrE. Figure adapted from Tomaz, R.⁽²²⁸⁾

ESCs and EpiSCs respectively represent a more naïve state of pluripotency and one primed for differentiation.^(6,7) While both pluripotent, they display distinct properties and behaviours. ESCs grow as small, dense, domed colonies, whereas EpiSC colonies form flatter monolayers, and tend to spontaneously differentiate at the edge of colonies. In ESC culture one maintains self-renewal with a combination of Leukaemia Inhibitory Factor (LIF) and serum-derived Bone Morphogenetic Proteins (BMPs), which stimulate the STAT3 and SMAD

signalling pathways respectively(8,9) One cannot culture EpiSCs under these conditions.(10,11) Instead they rely on Activin/Nodal and Fibroblast Growth Factor (FGF) signalling to maintain their self-renewal abilities (figure 1.4).(10,11) In fact the cell signalling pathways underlying ESC pluripotency can inhibit EpiSC pluripotency. For example, in EpiSCs, inhibition of STAT3 promotes self-renewal(10), and BMP4 exposure induces extra-embryonic differentiation.(11) ESCs and EpiSCs maintain distinct yet overlapping transcriptomes. Both cell types express *Pou5f1* (OCT4), *Sox2* and *Nanog*, (10,11) genes identified in ESCs as the foundation of pluripotency. Key ESC-specific (naïve) genes lose expression in EpiSCs (e.g. *Esrrb*; *Dppa3*, *Klf4*, *Tbx3* and *Zfp42*(10)), while genes associated with primed pluripotency gain expression (e.g. *Fgf5* and *Nodal*(11); *Otx2*(10)) (figure 1.2B). One can interconvert between ESCs and EpiSCs in culture by exposing them to each other's typical culture conditions.

1.1.4 Pluripotent stem cell heterogeneity and plasticity

In culture, ESCs appear both highly heterogeneous and plastic.(12) Experiments analysing the naïve pluripotency regulator *Zfp42* (REX1) exemplify these properties. Due to its high expression in ESCs and rapid downregulation in EpiSCs,(13) researchers often employ REX1 as a barometer of pluripotency. One can divide cultured ESCs into REX1 high expressing (REX1⁺) and REX1 low expressing (REX1⁻) subpopulations (figure 1.4).(13) mRNA expression of the naïve pluripotency associated proteins *Dppa3*, *Klf4*, and *Esrrb*, correlate with the expression of REX1. Conversely, transcription of the early differentiation markers *Fgf5*, *T* (BRACHYURY), and *Eomes* negatively correlates with REX1 expression. When isolated and seeded into culture both the REX1⁺ and REX1⁻ cell populations can regenerate the other. REX1⁺ cells readily contribute to chimeric embryos, whereas REX1⁻ cells do not, implying that ESC pluripotency favours a REX1⁺ environment. These experiments evidence that ESCs maintain distinct cell subpopulations in culture, and freely convert between these subpopulations. Furthermore these subpopulations exhibit different capacities for pluripotency (figure 1.2A).

In addition to the traditional serum/BMP4 conditions, one can culture mouse ESCs in the presence of two small molecule inhibitors targeting GSK3 and ERK, termed 2i.(14) In contrast to serum/BMP4 media, 2i supports a ground-state of pluripotency, which robustly impedes differentiation triggers.(15) Phenotypically, 2i culture conditions result in a much more homogenous cell population, compared to ESCs grown with serum/BMP4 (figure 1.2A).(12)

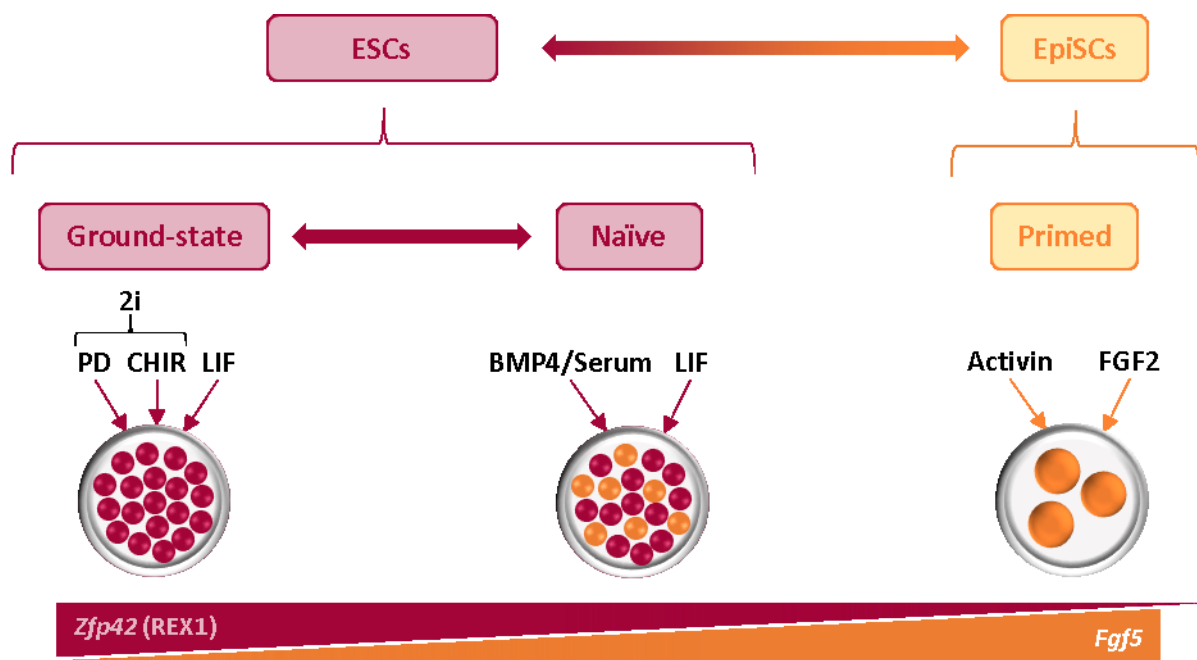


Figure 1.4 The heterogeneity of in vitro pluripotency.

The derivation and characterisation of EpiSCs showed that pluripotency can exist in distinct state in vitro: the naïve pluripotency of ESCs grown with BMP4/serum and LIF and the primed pluripotency of EpiSCs grown in the presence of Activin and FGF2. ESCs express higher levels of naïve (e.g. Zfp42) compared to primed (e.g. Fgf5) pluripotency transcription factors, and EpiSCs show the opposite pattern of expression. In addition to naïve pluripotency, the development of 2i culture conditions led to the discovery of the ground-state of pluripotency. Compared to naïve ESCs, ground-state ESCs appear much more homogenous in their expression of naïve pluripotency transcription factors.

1.4 Transcriptional regulation of pluripotency

Transcription factors bind to DNA to influence gene expression through myriad complex and context-dependent mechanisms of action. They do so by co-operating with other transcription

factors and co-regulators at genetic regulatory elements. The ESC transcriptional network presents one of the most extensively studied.

1.1.5 The OCT4-SOX2-NANOG triad

The OCT4-SOX2-NANOG (OSN) triad of transcription factors act as core components of the pluripotent transcriptional network. They work together to auto-regulate their own expression, as well as that of other pluripotency drivers.(16) OSN often co-localise to the same regulatory genetic elements, associated with both pluripotency and lineage-specific genes.(17,18) Perturbing the normal equilibrium of these proteins prompts differentiation.(19,20) Of these three proteins, researchers first identified OCT4 as critical to ESC pluripotency.(21–23) Highly expressed in the ICM and epiblast, the expression of its gene *Pou5f1* rapidly decreases upon differentiation.(24) Embryos carrying a homozygous deletion of *Pou5f1* cannot form pluripotent ICM and arrest their development at the blastocyst.(24) Depletion of OCT4 triggers differentiation towards TE.(25) Like OCT4, SOX2 shows high expression in the ICM and epiblast.(26) SOX2 also proves critical to ESC pluripotency.(27) In contrast to OCT4, ExE cells also express SOX2.(26) Over- and under-expression of SOX2 respectively triggers neuroectodermal(28) and TE(29) differentiation. Similarly to OCT4, SOX2 depletion compromises ICM formation resulting in embryonic lethality.(26) Researchers identified NANOG as a pluripotency transcriptional factor at approximately the same time as SOX2.(30) Mirroring OCT4 and SOX2, NANOG exhibits high expression in the ICM and epiblast,(31) and its depletion proves embryonically lethal. However, in stark contrast to OCT4 and SOX2, ESCs bearing homozygous deletions of *Nanog* can propagate themselves in culture. This indicates that while OCT4 and SOX2 prove absolutely necessary for self-renewal, NANOG is expendable.(32) Together OSN operate to maintain the pluripotency transcriptional network, but ESCs and EpiSCs contain other transcription factors that also contribute to pluripotency.

1.1.6 Naïve and primed pluripotency transcription factors

ESC pluripotency requires an additional cohort of naïve pluripotency factors, including: KLF2, KLF4, KLF5, ESRRB, TFCP2L1, TBX3, and GBX2.(33–36) ESCs express these proteins at

high abundance, whereas EpiSCs show minimal expression. During the transition to the primed EpiSC pluripotency, the expression of naïve transcription factors downregulates. Concurrently, OCT4 redistributes to new regulatory genetic elements where it interacts with primed pluripotency factors such as OTX2 and ZIC2/3 to establish the primed pluripotency transcriptional network.(34) Some of these primed pluripotency OCT4 binding sites contain low levels of OCT4 in the naïve state also. Over-expression of KLF2 and PRDM14,(37,38) and KLF4(33,39) in EpiSCs cultured in the presence of LIF, induces reprogramming to the naïve ESC state.

1.5 Epigenetic regulation of pluripotency

Mammalian cells can control gene expression at the DNA level without altering its sequence. We term this process epigenetics, from the Greek 'epi-' meaning 'upon', 'over', or 'above', co-opting a word originally coined by Charles Waddington in 1942(40) to describe the mechanisms governing the relationship between genetics and phenotype. Waddington believed this relationship to be "from a wider biological point of view, of crucial importance, since it is the kernel of the whole problem of development".(40) Today, epigenetics describes a vast and contentious array of cellular processes. Here I outline those processes most relevant to this research, and how they apply to pluripotency.

1.1.7 Chromatin organisation

Human and mouse DNA stretches to ~2 M(41,42) and ~1.4 M(43) in length. The cell must accommodate this sequence into a nucleus only 5-10 µM in diameter. In order to do so, the DNA wraps 1.65 times in 147 bp chunks around histone proteins to form a nucleosome. The histone proteins comprise an octet of two sets of four histone proteins: H2A, H2B, H3, and H4.(44) A section of linker DNA, joins each nucleosome forming a structure often compared to 'beads-on-a-string'. The histone H1 associates with linker DNA.(45) The nucleosomes pack into a chromatin fibre, which further organise to form the metaphase chromosome. Nucleosome packing varies in density. We call condensed chromatin heterochromatin and decondensed chromatin euchromatin. Under the light microscope, large areas of

heterochromatin appear visible as dark spots within the nucleus. Chromatin density impacts its accessibility to proteins, which in turn regulates gene expression. For example, if a transcription factor cannot access DNA it cannot promote expression of its target gene. Therefore, epigenetic mechanisms regulating chromatin accessibility play a vital role in controlling the transcriptional network of a cell.

1.1.8 The molecular basis of epigenetics

1.1.8.1 DNA methylation

DNA methylation is one of the longest studied and best characterised forms of epigenetics, and closely relates to chromatin accessibility and transcriptional regulation. DNA methyltransferases (DNMTs) catalyse the addition of methyl moieties, donated by S-adenosyl methionine (SAM), to one of the six atoms in the carbon ring of nucleotide bases. Typically, DNMTs act on cytosines followed by guanines (CpGs), however an increasing body of research evidences that they do also act on other dinucleotides,⁽⁴⁶⁾ particularly in ESCs.⁽⁴⁷⁾ This is commonly referred to as non-CpG methylation, reflecting the dominance of CpG methylation both biologically and in the literature. This work focuses on CpG methylation. CpG dinucleotides occur at lower than expected frequency throughout the genome,^(41–43,48,49) due to the increased mutability of methylated cytosines,^(50–52) which readily deaminate to yield thymine nucleotides.⁽⁵³⁾ Some regions of the genome contain greater than expected CpGs, termed CpG Islands (CGIs).⁽⁵⁴⁾ CpGs in these regions often appear hypomethylated relative to the rest of the genome when associated with active gene promoters. Some evidence supports that a hypomethylated state in the early embryo,⁽⁵⁵⁾ in combination with strong evolutionary selection⁽⁵⁶⁾ preserves the CpG content of these regions. The strong evolutionary selection of CpGs at CGIs evidences an important role in transcriptional regulation.

When DNMTs act on CpGs, the fifth atom of the cytosine carbon ring receives the methyl group, therefore we call this modified nucleotide 5-methylcytosine (5-mC or mCpG). The *de novo* DNMTs, DNMT3A and DNMT3B, add methyl groups to previously unmethylated

cytosines. Cells maintain CpG methylation through DNMT1 and its partner UHRF1, which recruits DNMT1 to hemi-methylated DN, to sustain methylation patterns after DNA replication.(57–59) DNA demethylation results from either a passive or enzyme-driven active processes. As an example of passive demethylation, the lack of DNMT1 and UHRF1, either through cellular depletion or re-localisation to the cytoplasm, would lead to demethylation through DNA replication. The TET family of proteins, TET1, TET2, and TET3, mediate active demethylation. They catalyse the oxidation of 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC), (60,61) 5-formylcystosine (5-fC), and 5-carboxylcytosine (5-caC).(62,63) TDG then excises 5-fC and 5-caC, which following base excision repair returns to cytosine.(62,64,65)

Researchers have implicated CpG methylation in numerous processes including: repressing transcription from retrotransposons; sex-specific monoallelic transcription of imprinted genes; X chromosome inactivation in female cells; directing transcription of mRNA alternative splicing; and regulating the chromatin accessibility of gene promoters and enhancers.(66) This study focuses on the role of DNA methylation at promoters and enhancers. CpG methylation of promoters(67–73) and enhancers(74–80) associates with chromatin condensation and transcriptional repression. Until recently these observations were purely correlative, but the advent of CRISPR-Cas9(81–88) technology has enabled the study of targeted DNA methylation. The CRISPR-Cas9 system employs a guide RNA (gRNA), which directs the complex to a specific genomic locus. The fusion to this system of the DNMT3A catalytic domain allows the targeted methylation of individual CpGs.(89–92) These studies evidence that transcriptional repression following promoter CpG methylation. However as Frank Lyko points out in his recent review,(93) these studies report only moderate effects on transcription, suggesting that CpG methylation represents one of a number of mechanisms required for complete and sustained transcriptional repression.

While recent work evidences that 5-hmC interacts with numerous transcription factors in different cell types including ESCs.(94–99), thus far the dissection of how CpG methylation

impacts transcription has largely focused on 5-mC. 5-mC broadly acts through two methods: altering the affinity of transcription factors to their DNA binding motifs, and mediating histone modifications through Methyl Binding Domain (MBD) proteins. CTCF is a classic example of the former. CpG methylation of the CTCF binding motif at the imprinted H19/Igf2 locus disrupts its interaction with DNA.(100,101) More recent work shows that, CpG methylation can both positively and negatively impact DNA-protein interactions.(102) Some transcription factors display highly context specific responses to CpG methylation. For example, the key naïve pluripotency transcription factor KLF4 shows a stronger and weaker association with two different versions of its DNA binding motif when the CpG it contains is methylated.(103,104) Some whole families of transcription factors display consistent responses to CpG methylation. For example, ~50% of nuclear receptors, ~60% of homeodomain proteins, and ~80% of POU proteins (including OCT4) increase in binding in response to CpG methylation of variants of their DNA binding motifs, whereas 100% of AP2 proteins, ~40% of bHLH proteins, and ~60% of bZIP proteins decrease in binding.(102) Thus methylation of even individual CpGs may positively or negatively impact transcription factor binding to promoters and enhancers.

1.1.8.2 Post-translational modifications of histone proteins

As discussed in chapter 1.5.1, nucleosome density inhibits transcription by preventing transcription factors from accessing DNA. Cells can modulate chromatin compaction by applying post-translational modifications to histone proteins. These modifications predominantly occur at the N-terminal tail, which extends from the globular body of the histone. Although these modifications include ADP ribosylation, butyrylation, deamination, phosphorylation, propionylation, sumoylation, and ubiquitinylation, lysine acetylation and methylation are the most well-characterised and relevant to this thesis.(105) Histone Acetyltransferases (HATs) catalyse the addition of acetyl groups, donated by acetyl CoA, to specific amino acids within histone proteins. Lysine acetylation nullifies its positive charge, weakening histone-DNA interactions.(106) HATs divide into five classes: GNAT, MYST, TAFII250, P300/CBP, and nuclear receptor co-activators.(107) Typically, HATs act within

large, multi-protein complexes.(106) Histone Deacetylases (HDACs) act in opposition to HATs, removing acetyl groups from histone proteins. HDACs also work within large, multi-protein complexes. For example, HDAC1 and HDAC2 form part of the NuRD complex. Generally, HATs decondense chromatin, whereas HDACs condense chromatin. Histone Lysine Methyltransferases (HKMTs) catalyse the addition of 1-3 methyl groups to lysine residues. The SET domain of HKMTs transfers a methyl group donated by SAM. Unlike acetylation, methylated lysines retain their positive charge. Compared to HATs and HDACs, which act on numerous residues, HKMTs behave more specifically, often only acting on a single residue and adding a specific number of methyl groups.(106)

Histone modifications co-ordinate with CpG methylation through MBD proteins. Researchers have identified 11 MBD proteins thus far: Methyl-CpG-binding protein 2 (MeCP2), MBDs 1-6, SETDB1-2, BAZ2A and BAZ2B.(108) In addition to their MBD, most of these proteins contain a transcriptional repression domain (TRD) through which they interface with other proteins.(109–111) For example, MeCP2 interacts with HDACs and HKMTs, and contributes to chromatin condensation.(112–114) Most MBDs preferentially bind to 5-mC over C.(115–117) Notably, MBD3 does not show a preference for 5-mC, but can bind 5-hmC.(118) MBD3 represses transcription as part of the NuRD complex, and is required for both the complexes formation and stability.(110,119) MBD proteins preferentially associate with genomic regions that typically contain above-average CpG methylation, however they appear depleted at repetitive DNA, which commonly appear hypermethylated.(120) MBD3 shows enrichment at H3K4me3 and H3K27ac marked site, which typically define active promoters and enhancers.(120) Furthermore, MBD3 can direct HDAC and DNMT recruitment through the NuRD complex.(121) In summary, together histone modifications and CpG methylation co-operate to epigenetically regulate transcription.

1.1.9 Epigenetic regulation of pluripotency

Epigenetic mechanisms play a critical role in regulating pluripotency both *in vivo* and *in vitro*. The embryonic lethality of homozygous deletion of *Mbd3* presents a clear example of the

fundamental necessity of epigenetic regulation to pluripotency. *Mbd3* null embryos die at approximately E3.5, coinciding with the onset of pluripotency.(122) The early embryo undergoes drastic re-organisation of its epigenome. Between conception and the early blastocyst, the cells severely deplete CpG methylation. Immediately prior to implantation, DNMT3A and DNMT3B rapidly lay down *de novo* CpG methylation(123–125) more predominantly in the ICM relative to the TE.(124) Subsequently, minimal *de novo* methylation occurs. Differences in histone modification profiles are also apparent. For example, the permissive H3K9ac and repressive H3K27me3 show higher levels in the ICM relative to the TE.(126) In ICM and ESCs, some genes display both the permissive H3K4me3 and the repressive H3K27me3, termed bivalency.(127,128) These genes remain silent until activated later in development by removal of H3K27me3.

Re-organisation of the epigenome accompanies the shift in transcriptome and cell signalling pathway dependency during pluripotency maturation (figure 1.5).(5,10,11,129,130) Within ground-state ESCs (2i+LIF), CpGs appear widely unmethylated.(131) This coincides with low expression of the *de novo* DNMTs DNMT3A and DNMT3B.(131) In tandem, they express high levels of TET1 and TET2, which increases 5-hmC content.(132,133) ICM cells at E3.5 show similarly low levels of CpG methylation, leading researchers to draw comparisons to ground-state ESCs. Many of the bivalently marked genes in naïve (serum+LIF) ESCs lack H3K27me3 in ground-state pluripotency without upregulating their transcription.(134) Naïve ESCs (serum+LIF) show higher genome-wide levels of CpG methylation, expression of DNMT3A and DNMT3B and lower expression of TET1 and TET2. Genome-wide CpG methylation further increases in EpiSCs (135,136) as does expression of DNMT3B.(137)

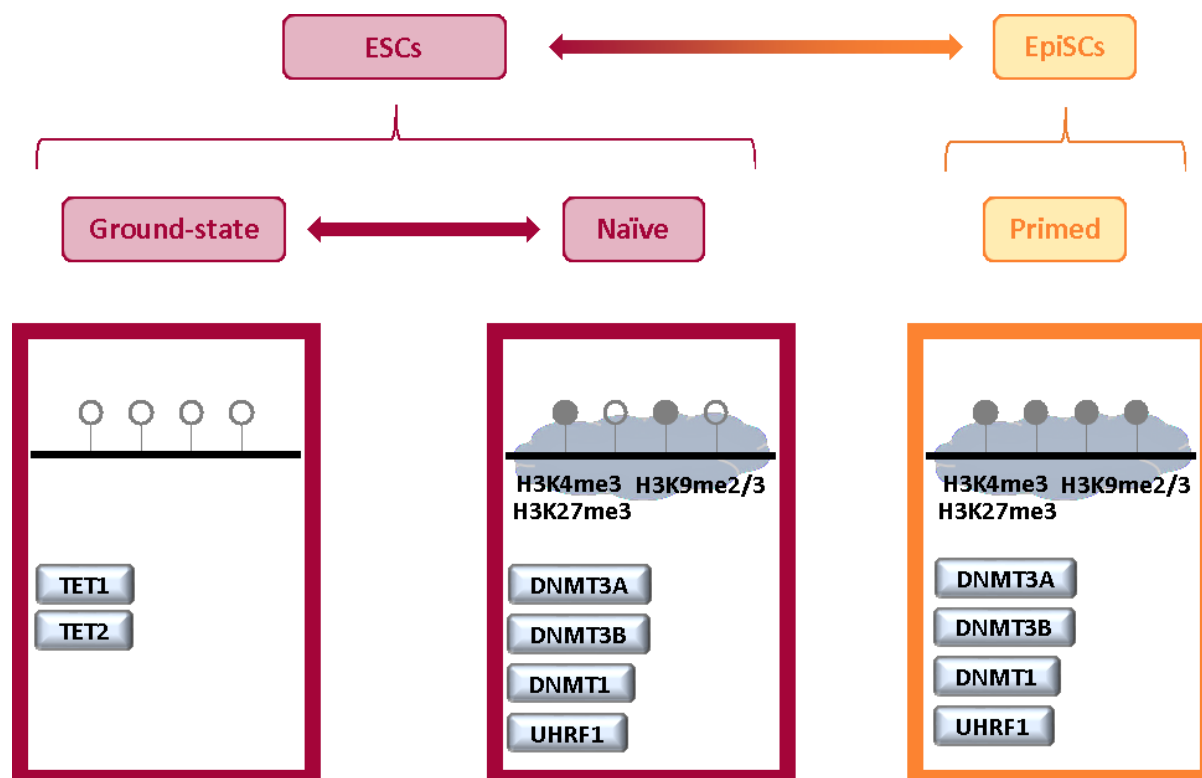


Figure 1.5 Epigenetic regulation of pluripotency.

The different stages of pluripotency display distinct epigenetic profiles. Ground-state ESCs globally show almost completely unmethylated CpGs (white lollipops) of their DNA (black line), due to high TET1 and TET2 expression. Naïve ESCs globally show intermediate levels of CpG methylation (grey lollipops), due to increased expression DNMTs. They also show increased bivalent and repressive histone modifications (grey cloud). EpiSCs further increase their CpG methylation through even higher expression of DNMTs.

1.6 Enhancers

1.1.10 The early history of enhancer research

In December 1981 Julian Banerji, Sandro Rusconi and Walter Schaffner published a paper characterising DNA elements within the Simian vacuolating virus 40 (SV40) genome.⁽⁶⁸⁾ They expressed SV40 carrying rabbit β -globin gene in HeLa cells and found a 200-fold increase in its expression.⁽⁶⁸⁾ Transcription depended on two 72 bp sequence motifs 200 bp upstream of the viral origin of replication (analogous to the eukaryotic transcription start site).⁽⁶⁸⁾ Researchers previously identified this region as necessary for transcription, but either

classified it as part of the promoter(138) or refrained from mechanistic speculation.(139) Importantly, Banerji *et al.*, made three key observations about how these regions drive transcription: they can act when positioned both upstream and downstream of a gene; they can act independently of orientation; they can influence gene promoters kilobases away.(68) These features distinguished the regions from promoters, which reside exclusively at the 5'-end of a gene and close to the transcription start site. They concluded that “these studies define a class of DNA elements with a mode of action that has not been heretofore described” and named the regions ‘enhancers’.(68) (See Schaffner’s fantastic 2015 autobiographical account for further detail and context.)(140)

The following year researchers added a further layer of complexity: enhancers can act in a cell-specific manner. In 1983, three back-to-back *Cell* publications reported a lymphocyte-specific enhancer downstream of an immunoglobulin gene.(141–143) Researchers began to discover cell-specific enhancers driving genes of diverse function in many metazoan cell types.(144–164) Enhancer research entered a phase of exponential growth that plateaued in approximately 1994 (figure 1.6). Then, in approximately 2008 our interest reignited, likely as a result of the plummeting cost of DNA sequencing technologies (figure 1.6B).(165) Indeed, PubMed indexed publications since 2007 containing in their title or abstract the words ‘gene’ and ‘enhancer’ and ‘ChIP’ and ‘sequencing’/‘seq’ correlate near perfectly (figure 1.6A) (Pearson correlation co-efficient = 0.96, in contrast the correlation with ‘gene’ and ‘promoters’ = -0.35). The renaissance in enhancer research has greatly improved our understanding of the capacity and mechanisms of enhancer function.

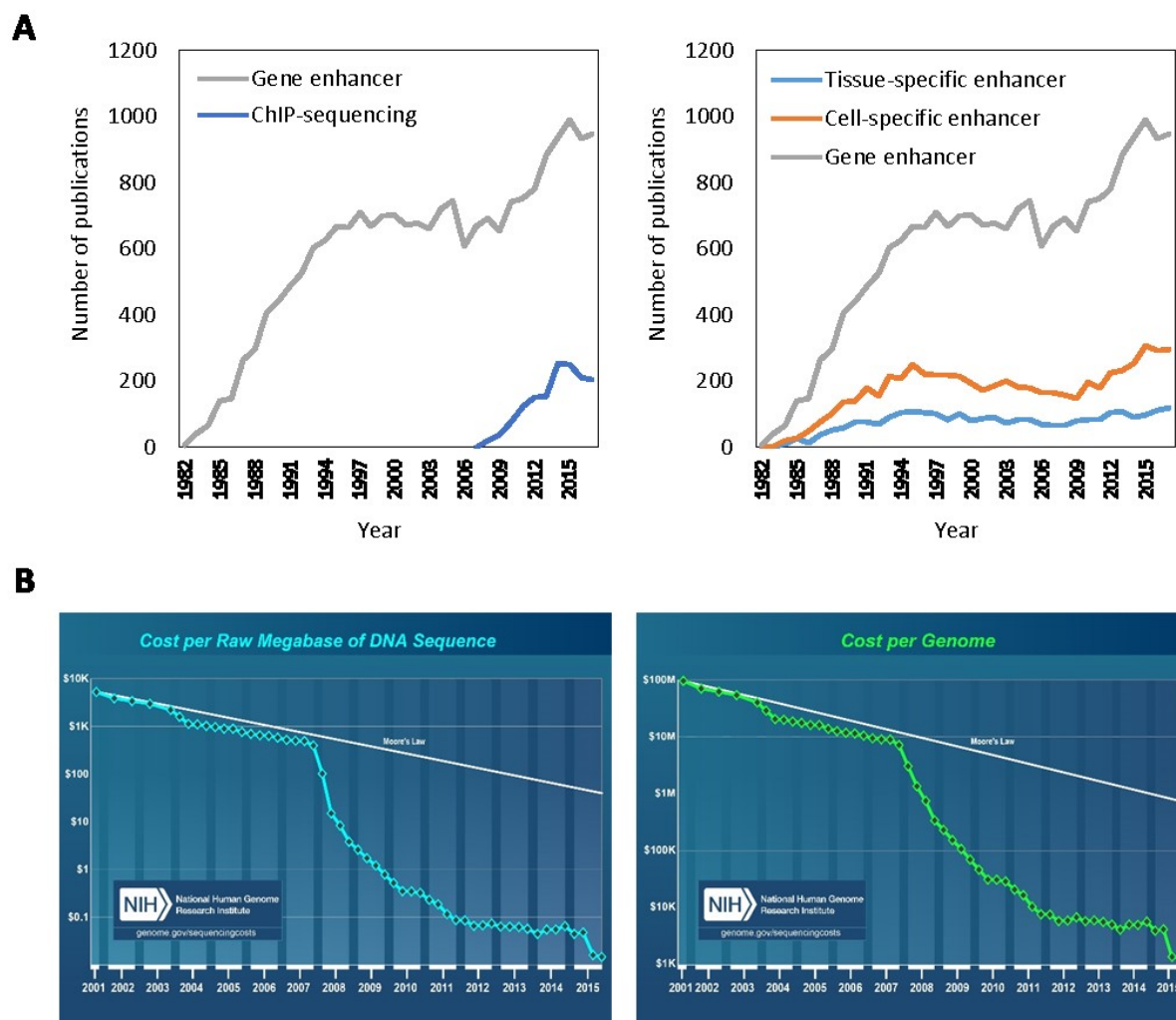


Figure 1.6 Enhancer research over time.

A. The number of PubMed and bioRxiv indexed publications pertaining to enhancers since 1982. I searched for publications containing in their title and/or abstract exact matches to the words indicated by the key in the top left-hand corner. **B.** The cost of DNA sequencing technology per raw megabase of sequence and per genome since 2001.(165) The resurgence in enhancer-related publications that occurred in ~2008 coincides with the massive decrease in the cost of DNA sequencing. Indeed, the number of ChIP-seq related papers published after 2008, strongly positively correlates with the number of enhancer related papers. This suggests that the boom in enhancer research resulted from the increasing accessibility of DNA sequencing.

1.1.11 The capacity and mechanisms of enhancer function

Enhancers contain high densities of transcription factor DNA binding motifs and recruit them in dense clusters.(18) The necessity of multiple transcription factors at enhancers stems from

their intrinsic genetic bias towards high nucleosome occupancy, related to their high GC content.(166) Nucleosome occupancy impedes access to DNA. The literature documents numerous instances where transcription factors work co-operatively, either with or without direct physical interaction, to overcome the energy threshold to displace these nucleosomes, thus unravelling the DNA and enabling protein-DNA interactions.(167,168) In developmental cell types, pioneer factors bind to nucleosomes to promote their displacement without needing to co-operate with other transcription factors.(169) In ESCs and EpiSCs OCT4 acts as a pioneer factor.(169,170) Currently there exist two dominant models of enhancer mechanism: the 'enhanceosome' and the 'billboard'. The enhanceosome model, first described by Thanos and Maniatis in 1995, posits that enhancer DNA acts as a scaffold for the binding of highly ordered and co-operative series of proteins. As such the DNA binding motifs resist variation in density, spacing, and orientation.(167,168,171) Computational modelling of protein-DNA binding suggests that the enhanceosome model may exist widely throughout the genome.(172) The billboard model, proposed by Kulkarni and Arnosti in 2003, requires a specific set of transcription factors, but affords greater flexibility to their DNA binding motif organisation.(173,174) This model allows for smaller sub-clusters of transcription factors working towards the same end, which in turn offers greater redundancy and reduced sensitivity to mutation.

Once the necessary transcription factors bind to DNA, they recruit co-activators to drive transcription. Co-activators may act by remodelling chromatin or facilitating interaction with the basal transcriptional machinery at gene promoters.(175,176) CRISPR-Cas9 targeting HAT domain of the co-activator P300 to candidate promoters and enhancers proves sufficient to boost transcription.(84) A complementary experiment targeting the HDAC LSD1 to enhancers inhibits transcription.(177) The versatile Mediator complex acts as a co-activator in all eukaryotic organisms.(178) In mammalian cell types it localises to active enhancers and promoters,(179,180) contributing to the formation Promoter Initiation Complex (PIC), nudging the PIC from the promoter to the downstream pause site (Promoter Escape), and releasing it

from its paused state (Pause Release).(178) Co-activators therefore represent the workhorses of transcriptional regulation.

1.1.12 The topological behaviour of enhancers

The advent of Chromatin Conformation Capture technology combined with DNA sequencing (e.g. HiC-seq) enables the detection of chromatin interactions. These tools revealed that chromatin interactions occur within self-contained chromosomal partitions termed Topological Associated Domains (TADs).(181,182) The enclosed nature of TADs segregates a chromosome into multiple local networks of genetic regulatory elements and their promoters, with its own epigenetic 'microclimate'.(181,182) CTCF and the cohesin complex, comprised of SMC1, SMC3, and RAD21, bind at high density to TAD boundaries, leading to the proposal that they contribute to TAD formation and insulation.(181,183,184) Several recent publications propose a 'loop extrusion model' of TAD formation, which relies on the co-operation of CTCF and cohesin.(185–187) Under the loop extrusion model a circlet of Loop Extrusion Factor (LEF) travels along a loop of DNA until it encounters a Boundary Element (BE) at which point the loop stabilises forming a TAD. Cohesin acts as a LEF and CTCF acts as a BE, possibly through homodimerisation. The directionality of CTCF binding governs how it forms protein-protein interactions,(188) so the two CTCFs must face in a complementary direction in order to dimerise. In summary, the development of HiC-seq and related technologies has led to substantial progress in our understanding of the topology and mechanisms underlying chromatin organisation at this resolution.

Despite the resulting advances in our understanding of chromatin organisation, some criticise current HiC-seq as able to detect only the most stable and frequent interactions, eliminating “most enhancers and promoters active in a given cell type”.(168) Furthermore, work in *Drosophila* indicates that enhancer-promoter loops pre-form during development and arrest in a transcriptionally paused position until the relevant cellular context arises,(189) thus an enhancer-promoter interaction in isolation does not evidence an active enhancer, let alone an active enhancer-promoter unit. In humans the promoters with which an enhancer interacts can

lie up to one megabase away,(190,191) a range three orders of magnitude larger than Banerji *et al.*,(68) observed. Multiple enhancers interact with the same promoters in *cis* and in *trans*.(189,190) In fact, almost all enhancers in the developing *Drosophila* embryo exhibit partial redundancy, which facilitates buffering against deleterious mutations.(192) Conversely, a single enhancer can interact with multiple promoters. In addition to driving a contextually relevant gene, tissue-specific enhancers appear to flexibly regulate large swathes of the genome in the developing mouse embryo.(193) Enhancers also promote the activity of other regulatory elements through enhancer-enhancer loops.(189,190) Seemingly all logical formulations of enhancer and promoter relationship exist in some manner in eukaryotes.

1.1.13 Super Enhancers and Typical Enhancers

In 2013 Whyte *et al.*,(179) reported the existence of “unusual enhancer domains” in mouse ESCs. Using the Mediator complex member MED1, they divided enhancers into high and low Mediator occupancy (figure 1.7A). Of the ~9,000 enhancers they identified, 231 accounted for 40% of the total MED1 signal. They respectively termed the high and low MED1 occupied enhancers Super Enhancers and Typical Enhancers respectively.(179,194,195) In addition to Mediator occupancy, Super Enhancers outstrip Typical Enhancers in terms of their: binding density of the ESC pluripotency-associated transcription factors and co-factors; chromatin accessibility as measured by DNase I hypersensitivity; density of the definitive enhancer histone modifications H3K27ac and H3K4me1; and size (median length: Super Enhancers = 8,667 bp, Typical Enhancers = 703 bp). Their composition of multiple stitched constituent enhancers in part explains their difference in size. Interestingly, ESC Super Enhancers and Typical Enhancers bind similar densities of OCT4, SOX2 and NANOG, however the naïve pluripotency transcription factors ESRRB and KLF4 exhibit preferential enrichment at Super Enhancers (figure 1.7B-C).(179) When Whyte *et al.*, assigned genes to Super Enhancers based on proximity they found almost all were key ESC identity regulators and few were house-keeping genes, implying that Super Enhancers direct the expression of cell-specific genes. ESCs transcribe Super Enhancer associated genes at higher levels than Typical

Enhancer associated genes (figure 1.7D-E). Only Super Enhancer associated genes showed consistent loss of expression upon knock down of *Med1* and *Pou5f1* (OCT4) in ESCs. In summary, Whyte *et al.*, presented Super Enhancers as auto-regulating cell-specific enhancers, but it was unclear if ESC Super Enhancers continued their activity, were decommissioned, or were otherwise reorganised upon differentiation to EpiSCs.

1.7 Hypothesis and aims

Hypothesis: Super Enhancers defined in ESCs represent only naïve pluripotency and decommission under the primed pluripotency of EpiSCs.

As described above, pluripotency requires a complex transcriptional network composed of many proteins working together to auto-regulate, promote the expression of pluripotency genes, and repress the expression of differentiation triggers. The transcriptional networks controlling ground-state, naïve and primed phases of pluripotency exhibit similar, but distinct properties. It follows that these stem cells maintain unique, but perhaps overlapping, webs of enhancers. Whyte *et al.*, defined ESC Super Enhancers in cells grown under serum+LIF conditions. If Super Enhancers represent cell-specific enhancers, EpiSCs likely decommission ESC Super Enhancers in favour of their own EpiSC Super Enhancers. If these assumptions prove correct, ESCs and EpiSCs present a model to study cell-specific enhancer re-organisation. Therefore I hypothesised that ESC Super Enhancers represent only naïve pluripotency, and decommission during the transition from naïve ESCs to the primed EpiSCs.

This thesis first aims to determine whether ESC Super Enhancers indeed decommission in EpiSCs. To do this I will employ publicly available and shared DNA sequencing data to examine epigenetic and transcriptional changes between ESCs and EpiSCs. Secondly, I will repurpose these data to identify potential mechanisms by which this decommissioning may occur, followed by wet-lab experimentation to extend and validate any findings.

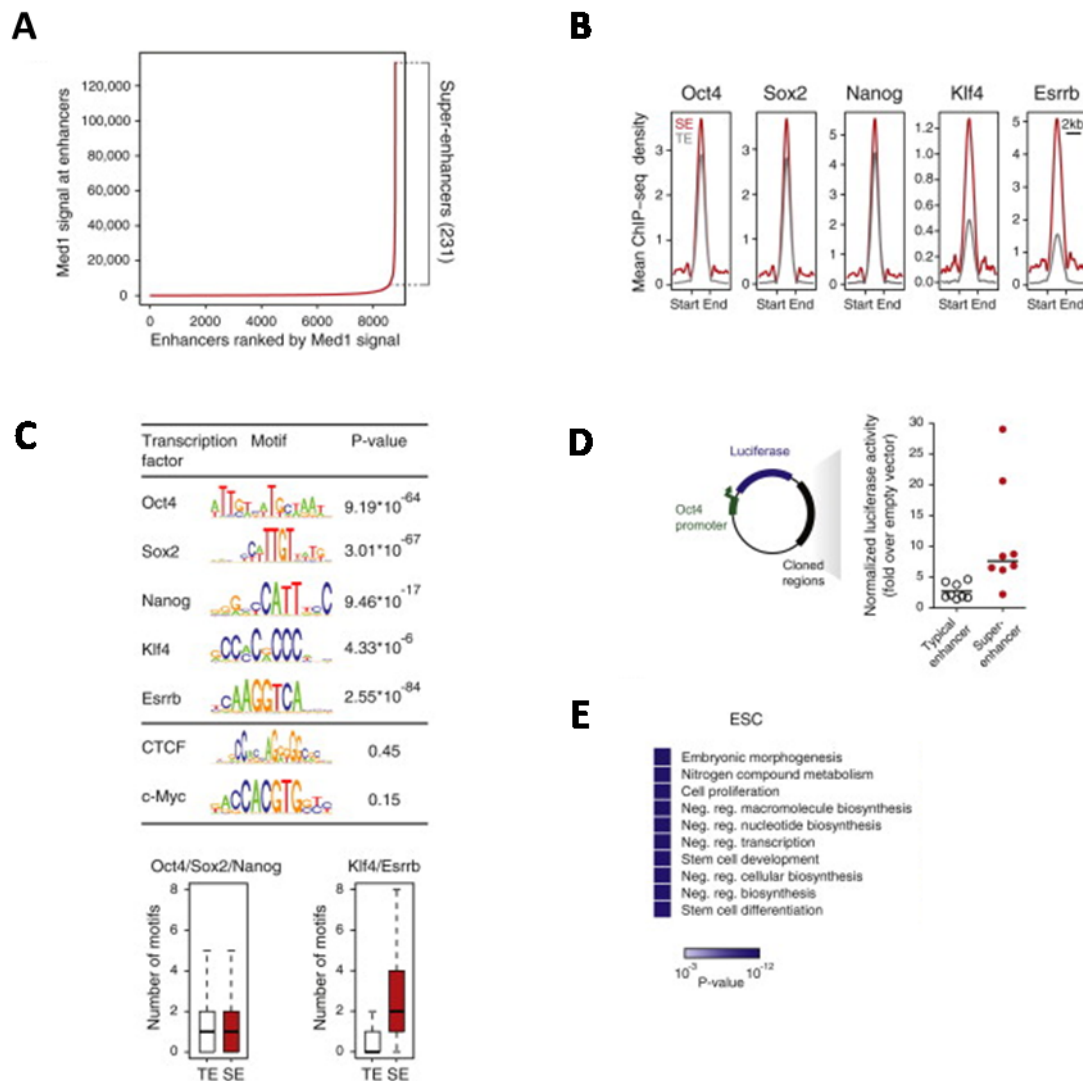


Figure 1.7 ESC Super Enhancers act as turbo-charged cell-specific enhancers.

A. Whyte et al., defined Super Enhancers by ranking and plotting MED1 binding density, using a slope > 1 as their cut-off. **B.** Super Enhancers compared to Typical Enhancers bind similar densities of OSN, but much higher densities of the naïve pluripotency transcription factors KLF4 and ESRRB. **C.** The motif enrichment analysis of these regions confirms that the Klf4 and Esrrb DNA binding motifs appear more frequently at Super Enhancers compared to Typical Enhancers. **D.** Whyte et al., used OCT4 promoter driven luciferase assays to test the expression of Super Enhancer and Typical Enhancer associated genes. They found that Super Enhancer associated genes showed greater luciferase activity than Typical Enhancer associated genes, indicating that Super Enhancers drive higher levels of transcription. **E.** Gene ontology analysis of Super Enhancer associated genes indicated that Super Enhancers drive the expression of embryonic development and stem cell factors. Figure adapted from Whyte et al. [179]

2 MATERIALS & METHODS

2.1 Materials

2.1.1 Buffers, solutions and gels

Alkaline phosphatase staining

AP fixation solution	25.5% (v/v) Citrate solution
	66.3% (v/v) Acetone
	8.2% (v/v) Formaldehyde 37%
AP staining solution	1.9% (v/v) Na-Nitrite
	1.9% (v/v) FRV-alkaline
	94.3% (v/v) dH ₂ O
	1.9% (v/v) Naphthol Alkaline

Genomic DNA extraction

Lysis Buffer	200 mM NaCl
	100 mM Tris-HCL pH 7.4
	5 mM EDTA
	0.2% (w/v) SDS
	0.5 mg/mL Proteinase K

ChIP Protocol A

Lysis Buffer	25 mM HEPES pH 7.9
	1.5 mM MgCl ₂
	10 mM KCl
	0.1% (v/v) NP-40
IP Buffer	50 mM HEPES pH 7.9
	140 mM NaCl
	1 mM EDTA
	1% (v/v) Triton X-100
	0.1% (w/v) Na-Doc
Wash Buffer 1	0.1% (w/v) SDS
	50 mM HEPES pH 7.9
	500 mM NaCl
	1 mM EDTA
	1% (v/v) Triton X-100
	0.1% (w/v) Na-Doc
	0.1% (w/v) SDS

	20 mM Tris pH 8
	1 mM EDTA
Wash Buffer 2	250 mM LiCl
	0.5% (v/v) NP-40
	0.5% (w/v) Na-Doc
TE Buffer	10 mM Tris pH 8
	1 mM EDTA pH 8
Elution Buffer	50 mM Tris pH 8
	1 mM EDTA pH 8
	1% (w/v) SDS

ChIP Protocol B

	50 mM HEPES pH 7.5
	140 mM NaCl
	1 mM EDTA
Lysis Buffer 1	10% (v/v) glycerol
	0.5% (v/v) NP-40
	0.25% (v/v) Triton X-100
	200 mM NaCl
	1 mM EDTA
Lysis Buffer 2	0.5 mM EGTA
	10 mM Tris pH8
	50 mM Tris pH 8
	10 mM EDTA
	150 mM NaCl
Lysis Buffer 3	0.1 % (w/v) SDS
	0.95 % (v/v) NP40
	0.1% (w/v) Na-Doc
	150 mM NaCl
	20 mM Tris pH 8
Wash Buffer 1	2 mM EDTA
	0.1% (w/v) SDS
	1% (v/v) TritonX-100
	2 mM EDTA
	20 mM Tris pH8
Wash Buffer 2	500 mM NaCl
	0.1% (w/v) SDS
	1% (v/v) TritonX-100

Wash Buffer 3	0.25 M LiCl
	1% (v/v) NP-40
	1 % (w/v) NaDoc
	1 mM EDTA
TE Buffer	10 mM Tris pH8
	50 mM Tris-HCL pH8
	1 mM EDTA pH8
	50 mM NaCl
Elution Buffer	50 mM Tris-HCL pH8
	1 mM EDTA pH8
	1% (w/v) SDS

ChIP Protocol C

Lysis Buffer	100 mM NaCl
	50 mM Tris-Cl pH8.1
	5 mM EDTA pH8.0
	0.02% (v/v) NaN ₃
	0.5% (w/v) SDS
IP Buffer	1 V Lysis Buffer + 0.5 V Dilution Buffer
Dilution Buffer	100 mM Tris-HCl pH8.6
	100 mM NaCl
	5 mM EDTA pH8.0
	0.2% (w/v) NaN ₃
	5% (v/v) Triton X-100
Wash Buffer 1	150 mM NaCl
	20 mM Tris-HCl pH8.1
	5 mM EDTA pH8.0
	5.2% (w/v) Sucrose
	0.02% (w/v) NaN ₃
	1% (v/v) Triton X-100
Wash Buffer 2	0.2% (w/v) SDS
	0.5% (w/v) DOC
	1 mM EDTA
	250 mM LiCl
	10 mM Tris-HCl pH8.0
	0.02% (w/v) NaN ₃
	0.5% (v/v) NP-40

Wash Buffer 3	0.1% (w/v) DOC
	1 mM EDTA
	50 mM HEPES pH7.5
	500 mM NaCl
	1% (v/v) Triton X-100
	0.02% (w/v) NaN ₃
TE Buffer	10 mM Tris-HCl pH 8
	1 mM EDTA
Elution Buffer	1% (w/v) SDS
	0.1 M NaHCO ₃

2.1.2 Chemicals, solvents and reagents

Reagent	Source
Acetone	Fisher Scientific
Acrylamide/Bis Solution	Biorad
Agarose	Sigma-Aldrich
All Trans Retinoic Acid	Sigma-Aldrich
Ammonium Persulfate (APS)	Sigma-Aldrich
Bovine Serum Albumin (BSA)	PAA
Deoxycholic acid (DOC)	Sigma-Aldrich
Deoxynucleotide Triphosphate (dNTPs) 10 mM	Lifetechnologies
DEPC-Treated Water	Ambion
Dimethyl Sulfoxide (DMSO)	Sigma-Aldrich
DL-Dithiothreitol (DTT)	Sigma-Aldrich
DL-Dithiothreitol (DTT) <i>for molecular biology</i>	Lifetechnologies
Dyalysis Tubing	ThermoScientific
Dynabeads Protein G-coupled	Lifetechnologies
Ethanol	Fisher Scientific
Ethidium Bromide	Sigma-Aldrich
Ethylene Diamine Tetraacetic Acid (EDTA)	Sigma-Aldrich
Ethylene Glycol Tetraacetic Acid (EGTA)	Sigma-Aldrich
FLAG coated sepharose beads	Sigma-Aldrich
FLAG tripeptide	Sigma-Aldrich
Formaldehyde Solution, 37%	Sigma-Aldrich
Glycerol	Sigma-Aldrich
Glycine	Sigma-Aldrich
Glycogen Blue (GlycoBlue) carrier	Ambion
Sulphuric acid (H ₂ SO ₄)	Sigma-Aldrich
HEPES	Sigma-Aldrich
Horse Serum	Sigma-Aldrich
Isopropanol	Fisher Scientific
Laemmli Sample Buffer	BioRad
LB	MP Biomedicals
Lipofectamine 2000 Transfection Reagent	Lifetechnologies
Lithium Chloride (LiCl)	Sigma-Aldrich
Magnesium Sulphate (MgSO ₂)	Novagen
Methanol	Fisher Scientific
MgCl ₂	Lifetechnologies
Nonidet P-40 (NP-40)	Sigma-Aldrich
Normal Goat Serum	Sigma-Aldrich
Oligonucleotides	Lifetechnologies
Paraformaldehyde (PFA)	Sigma-Aldrich
Phenol-Chloroform	Sigma-Aldrich
Phosphate-buffered Saline (PBS) capsules	Fisher

Potassium Chloride (KCL)	Sigma-Aldrich
Protein A-Sepharose	Sigma-Aldrich
Random Hexamers	Lifetechnologies
Sodium Azide (NaN ₃)	Sigma-Aldrich
Sodium Chloride (NaCl)	VWR
Sodium Deoxycholate (Na-Doc)	Sigma-Aldrich
Sodium Dodecyl Sulphate (SDS) Solution 10%	Fluka
Sodium Hydrogenocarbonate (NaHCO ₃)	Sigma-Aldrich
Spermidine	Sigma-Aldrich
Spermine tetrahydrochloride	Sigma-Aldrich
Sucrose	Sigma-Aldrich
Trichloroacetic acid (TCA)	Sigma-Aldrich
Tris Base	Sigma-Aldrich
Triton X-100	Sigma-Aldrich
Tween 20	Sigma-Aldrich
β-Mercaptoethanol	Sigma-Aldrich

2.1.3 Antibiotics

Antibiotic	Concentration (µg/mL)	Catalogue Number	Source
Geneticin (G418)	100	P11-012	PAA
Puromycin	1	P7255	Sigma-Aldrich

2.1.4 Antibodies

A dilution factor of 1:1,000 indicates 1 part antibody in 999 parts diluent.

2.1.4.1 Primary antibodies

Antibody	Host species	Catalogue number	Supplier	Western blot (dilution factor)	IF (dilution factor)	ChIP (µg)
Esrrb	Mouse	H6705	R&D Systems	1:1,000	...	...
FLAG M2	Mouse	F1804	Sigma	1:1,000	1:500	10
Gapdh	Mouse	sc-137179	SantaCruz	1:20,000	...	...
GFP	Rabbit	ab290	Abcam	...	1:1,000	...
H3	Rabbit	ab1791	Abcam	1:50,000	...	...
H3K27ac	Rabbit	ab4729	Abcam	1:100,000	...	5
Med1	Rabbit	A300-793A	Bethyl Labs	1:1,000	...	8
Nanog	Rabbit	RC-AB002PF	Reprocell	1:2,000	...	...
Ncoa3	Rabbit	sc-9119	SantaCruz	1:2,500	...	...
Oct4	Goat	sc-8628	SantaCruz	1:1,000	...	5
Oct4	Mouse	sc-5279	SantaCruz	...	1:200	...
p300	Rabbit	sc-585	SantaCruz	1:1,000	...	5
V5	Mouse	46-0705	Invitrogen	1:2,000	1:1,000	...
Vinculin	Mouse	V9269	Sigma	1:250,000	...	...
α-Tubulin	Mouse	T6074	Sigma-Aldrich	1:20,000	...	...

2.1.4.2 Secondary antibodies

Antibody	Catalogue number	Supplier	Western blot (dilution factor)	IF (dilution factor)
Alexa488 Goat Anti-Mouse IgG	A11034	ThermoFisher	...	1:500
Alexa488 Goat Anti-Rabbit IgG	A11029	ThermoFisher	...	1:500
Alexa488 Donkey Anti-Goat IgG	A11055	ThermoFisher	...	1:500
Alexa488 Goat Anti-Rat IgG	A11006	ThermoFisher	...	1:500
Alexa633 Goat Anti-Mouse IgG	A21052	ThermoFisher	...	1:500
Alexa633 Goat Anti-Rabbit IgG	A21071	ThermoFisher	...	1:500
Goat Anti-Mouse IgG-HRP	sc-2004	SantaCruz	1:2,000	...

Goat Anti-Rabbit IgG-HRP	sc-2005	SantaCruz	1:5,000	...
Donkey Anti-Goat IgG-HRP	sc-2020	SantaCruz	1:5,000	...

2.1.4.3 Standard IgG antibodies

Species	Catalogue Number	Supplier	ChIP (µg)
Mouse IgG	sc-2025	Santa Cruz	2.5
Rabbit IgG	sc-2027	Santa Cruz	2.5

2.1.5 Cell lines

Cell line	Parental cell line	Genetic background	Details
COS-7	COS-7	CV-1	A monkey renal cell line transformed with mutant SV40.(196)
E14	E14Tg2A	129/Ola	HGPRT deficient cells, derived from the first ESC lines cultured by ML Hooper in 1987. (197)
EpiSC 8.73	Embryo-derived	C57Bl/6xDBA/2	Embryo-derived EpiSCs.(198)
ESC ^{Esrrb} -/-	E14Tg2A	C57BL/6	E14Tg2A cells containing Sox2:Cre-ERT2 and loxP flanked Esrrb second exon transgenes, treated with tamoxifen to delete the Esrrb second exon DNA sequence.(199,200)
ESC ^{Esrrb} +/+	E14Tg2A	C57BL/6	E14Tg2A cells containing Sox2:Cre-ERT2 and loxP flanked Esrrb second exon transgenes.(199,200)
ESC ^{Nanog} -/-	RCNβH-B(t)	129/Ola	E14Tg2A cells containing a ROSA:Cre-ERT2 and loxP flanked Nanog transgenes, treated with tamoxifen to delete the Nanog DNA sequence.(201)
ESC ^{Nanog} +/-	RCN	129/Ola	E14Tg2A cells containing a ROSA:Cre-ERT2 and loxP flanked Nanog transgenes. (201)
ESC ^{Ncoa3} -/-		C57/129	Provide by Austin Cooney.
ESC ^{Ncoa3} +/-		C57/129	Provide by Austin Cooney.
MEF	Embryo-derived	CF-1	MEF derived from CF-1 mice.

2.1.6 Enzymes

Enzyme	Catalogue Number	Source
DNase I (RNase free)	79254	Qiagen
FastStart™ Taq DNA Polymerase	12032945001	Roche
Micrococcal nuclease	N3755	Sigma
Proteinase K	P2308	Sigma-Aldrich
RNase A (DNase free)	R6513	Sigma-Aldrich
RNase Out	10777-019	Lifetechnologies
Superscript II Reverse Transcriptase	18064014	Lifetechnologies
Superscript III Reverse Transcriptase	18080044	Lifetechnologies
SYBR Green Taq JumpStart ReadyMix	S4438	Sigma-Aldrich

2.1.7 Kits

Kit	Catalogue #	Source
Amersham ECL Prime Western Blotting Detection Reagent	25080060	GE Healthcare
EZ DNA Methylation-Gold™ Kit	D5006	Zymo Research
EZ-96 DNA Methylation-Gold™ Kit (Deep Well Format)	D5008	Zymo Research
Illustra GenomiPhi V2 DNA Amplification Kit	25-6600-30	GE Healthcare
Leukocyte Alkaline Phosphatase Kit	86R-1KT	Sigma-Aldrich
Pierce BCA Protein Assay	23225	Thermo Scientific
RNAeasy mini	74106	Qiagen
RNase-Free DNase Set	79254	Qiagen

2.1.8 Inhibitors

2.1.8.1 *Protease inhibitors*

Inhibitor	Catalogue #	Source	Concentration (dilution factor)
Phenylmethylsulfonyl fluoride (0.1M)	93482	Sigma-Aldrich	1:1,000
Complete Protease inhibitor tablets (50x)	11697498001	Roche	1:50

2.1.8.2 *Phosphatase inhibitors*

Inhibitor	Catalogue Number	Source	Mass (mM)
Sodium Fluoride (NaF)	201154	Sigma	5
Sodium Orthovanadate (Na ₃ VO ₄)	S6508	Sigma	2

2.1.8.3 *Signalling pathway inhibitors*

Inhibitor	Catalogue Number	Source	Mass (mM)
CHIR99021 - GSK3 inhibitor	130-103-926	Miltenyi Biotech	3
PD0325901 - MEK inhibitor	04-0006	Stemgent	1

2.1.9 Plasmids

Plasmid	Supplier
pSG5-Flag-mEsrrb	Vanacker et al(202)
pSG5-Flag-mEsrrb-Rescue	Vanacker et al(202)
pSG5-Flag-mEsrrb-AF-2 mut1	Vanacker et al(202)
pSG5-Flag-mEsrrb-AF-2 mut2	Vanacker et al(202)

2.1.10 Tissue culture components

2.1.10.1 *Media recipes*

2.1.10.1.1 **ESC media**

2i
1:1 Neurobasal and DMEM-F12
0.5% (v/v) N2
1% (v/v) B-27
0.1 mM 2-Mercaptoethanol
2 mM L-Glutamine
50 U/mL Penicillin
50 µg/mL Streptomycin
LIF ¹
3µM CHIR99021 ²
1µM PD0325901 ²

E14
GMEM-BHK21
10% (v/v) FBS
0.1 mM 2-Mercaptoethanol

¹ LIF was produced in-house. For each batch the optimal volume was determined by self-renewal assay.

0.24% (w/v) Sodium Bicarbonate 1 mM Sodium Pyruvate 0.1 mM Non-essential Amino Acids 2 mM L-Glutamine 50 U/mL Penicillin 50 µg/mL Streptomycin LIF ¹

2.1.10.1.2 EpiSC media

N2B27 (poor)

1:1 mixture of Neurobasal and DMEM-F12 0.5% (v/v) N2 1% (v/v) B-27 without vitamin A 0.1 mM 2-Mercaptoethanol 2 mM L-Glutamine 50 U/mL Penicillin 50 µg/mL Streptomycin 20 ng/mL Activin A ² 12 ng/mL bFGF ²
--

N2B27 (rich)

1:1 mixture of Neurobasal and DMEM-F12 1% (v/v) N2 2% (v/v) B-27 without vitamin A 0.1 mM 2-Mercaptoethanol 2 mM L-Glutamine 50 U/mL Penicillin 50 µg/mL Streptomycin 20 ng/mL Activin A ² 12 ng/mL bFGF ²
--

CDM-BSA

1:1 IMDM and Ham's F12 5 mg/mL BSA 1% (v/v) Concentrated lipids 15 µg/mL Transferrin 450 µM Monothioglycerol 7 µg/mL Insulin 20 ng/mL Activin A ² 12 ng/mL bFGF ²
--

2.1.10.1.3 Other

MEF/Cos-7 media

DMEM 10% (v/v) FBS 0.1 mM 2-Mercaptoethanol 2 mM L-Glutamine 25 U/mL Penicillin 25 µg/mL Streptomycin
--

Epiblast coating media

DMEM

² This reagent was added freshly to the media immediate before application to the cells.

10% (v/v) FBS 25 U/mL Penicillin 25 µg/mL Streptomycin
--

Cryopreservation media

Normal culture media 10% DMSO 50% FBS ³
--

2.1.10.2 Media supplements

Supplement	Amount (µg/mL)	Catalogue number	Source
Activin A	0.02	...	In house
bFgf	0.02	100-18B	Peptrotech
Transferrin	15	652202	Roche
LIF	... ⁴	...	In house
Insulin	7	1376497	Roche
Heparin	1	H3393	Sigma-Aldrich
N2 (100x)	...	17502-048	Lifetechnologies
B27 -vitamin A (50x)	...	12587-010	Lifetechnologies
B27 (50x)	...	17504-044	Lifetechnologies
Concentrated lipids	10	066-01905H	Lifetechnologies
BSA	5	BSA B Lot No: 68-1137, EQBAH68-1000	Europa
Monothioglycerol	40	M6145-100ML	Bioproducts
Penicillin-Streptomycin	10	15140122	Sigma
			Life technologies

2.1.10.3 Other cell culture reagents

Supplement	Catalogue number	Source
TrypLE™ Express Enzyme (1X), no phenol red	12604013	ThermoFisher

2.1.10.4 Culture vessel coatings

Coating	Cell lines
0.1-0.2% Gelatine	ESCs
FCS	EpiSCs
Fibronectin (10 µg/mL)	EpiSCs

2.1.11 Resources and websites

2.1.11.1 Computer software

Name	Authors
Affy	Gautier, et al, 2004(203)
Bedtools 2.22.1	Quinlan and Hall, 2010(204)
Bowtie2 2.2.3	Langmead and Salzberg, 2012(205,206)
FastQC v0.11.2	Andrews, 2010(207)
Fastq-dump 2.4.2	NCBI(208)

³ FCS was only included in the cryopreservation media if it was part of the cells' normal culture media.

⁴ For each batch the optimal volume was determined by self-renewal assay.

Filezilla client 3.25.1	FileZilla(209)
GenomicRanges	Lawrence, <i>et al</i> , 2013(210)
GNU Image Manipulation Program (GIMP) 2	The GIMP Team(211)
Homer v4.7.2	Heinz, <i>et al</i> , 2010(212)
Limma	Smyth GK, 2005(213)
Lumi	Du, <i>et al</i> , 2008(214)
MACS2 2.1.1	Zhang, <i>et al</i> , 2008(215)
Mendeley	The Mendeley Team(216)
Microsoft Office 2013	Microsoft(217)
Picard Tools 1.128	Broad Institute(218)
PuTTY 0.63	The PuTTY Team(219)
R 3.3.2	R Core Team(220)
Samtools 1.1	Li, <i>et al</i> , 2009(221)
ShortRead	Morgan, <i>et al</i> , 2009(222)
SRA Toolkit 2.4.2	NCBI(208)
Trimmomatic 0.33	Bolger, <i>et al</i> , 2014(223)
XML	Duncan Temple Lang and the CRAN Team, 2016(224)

2.1.11.2 Websites

Tool	URL
Bioconductor	https://www.bioconductor.org/
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Box	https://www.box.com/
Cistrome Data Browser	http://cistrome.org/db/
Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/
GitHub	https://github.com/
Primer-BLAST	https://www.ncbi.nlm.nih.gov/tools/primer-blast/
Pubmed	http://www.ncbi.nlm.nih.gov/pubmed/
UCSC Genome Browser	http://genome.ucsc.edu
WashU Epigenome Browser	http://epigenomegateway.wustl.edu/

2.2 Methods

2.2.1 Wet lab methods

2.2.1.1 Cell culture

All cell lines were incubated at 37°C with 5% CO₂. Media recipes are listed in section 2.1.10.1. Media was changed daily. Cells were not cultured beyond 45 passages. All manipulation took place in a laminar flow cabinet under aseptic conditions. Cell lines were regularly tested for mycoplasma contamination.

2.2.1.1.1 Embryonic stem cell culture

All ESC lines were cultured in without a feeder layer, except for the ESC^{Ncoa3^{-/-}} and ESC^{Ncoa3^{+/+}} cell lines. Upon thawing, these cells were established on a Mouse Embryonic Fibroblast (MEF) feeder layer, until regular growth was observed at which point the feeder layer was removed and the MEFs passaged out of the culture. All other ESC lines were grown in culture vessels pre-coated with 0.1% gelatine in PBS. ESC^{Esrrb^{-/-}} showed a defect in adherence and thus both it and its wild-type counterpart were grown on 0.2% gelatine in PBS.

By default, ESC lines were grown in the FBS-containing E14 media. To adapt cells to serum-free 2i media they were first seeded in E14 media on culture vessels coated with 0.2% gelatine in PBS. After 24 hours the cells were washed with PBS and 2i+LIF media was applied. Cells were adapted to this media for 7 days before any work was performed.

To passage ESCs, cells were incubated in TrypE Express for 2-3 minutes at room temperature. Cells were collected and centrifuged for 5 minutes at 1,000-1,200 RPM in the appropriate media. Cells were most commonly seeded into new culture vessels at a 1:5 ratio every other day, however these parameters could vary from 1:3-1:10 ratios and up to every 3 days.

2.2.1.1.2 **Epiblast stem cell culture**

EpiSC lines were grown in culture vessels coated with fibronectin or 0.1% gelatine and Epiblast Coating Media. The cells were cultured in either N2B27 or CDM-based media. Both media were supplemented with bFGF and Activin A immediately prior to application. To passage EpiSCs, cells were incubated in PBS at 37°C for 2 minutes to detach clumps of cells from the culture vessel. These cell clumps were collected and centrifuged for 5 minutes at 1000 RPM and then gently pipetted into smaller clumps. Cells were most commonly seeded into new culture vessels at a 1:5 ratio every 3 days, however these parameters could vary from 1:2-1:5 every 2-3 days.

2.2.1.1.3 **Cryopreservation**

All cell lines were stored long-term in liquid nitrogen within a self-pressurising cryogenic storage dewar. To prepare cells for cryopreservation, cells were collected and spun down in cryogenic tubes and supplied with cryopreservation media. These tubes were stored at -80°C for 1-30 days and then transferred to the storage dewar.

To thaw frozen cells, cryogenic tubes were incubated in a 37°C water bath for no longer than 2 minutes. The contents of the tubes were then diluted 1:10 with the appropriate media for that cell line. Cells were centrifuged for 5 minutes at 1000 RPM and then seeded according to their normal seeding protocol. To remove dead cells, media was changed 12 hours after seeding.

2.2.1.1.4 **Producing LIF**

LIF was produced in-house. COS-7 cells were cultured in their media for 1 passage and then expanded into 20 x 10 cm diameter petri dishes with 10 mL of E14 media without LIF. At 70% confluency, 24 µg of LIF-expressing vector was added to each petri dish. The media was collected after 24 and 48 hours, combined, filtered, aliquoted, and stored at -20°C. To determine the optimal working volume of each batch of LIF, the fresh batch was titrated and a range of concentrations, plus no LIF, was used in an alkaline phosphatase staining self-

renewal assay. The results were compared to the previous batch and commercially available LIF.

2.2.1.2 *Embryonic stem cell phenotype assays*

2.2.1.2.1 **Self-renewal assay**

To determine the colony formation efficiency of ESCs, 8,000 cells/well were seeded into a 6-well plate. The cells were grown standard culture media either with or without LIF. Media was replaced every day. After 8 days the plates were fixed and alkaline phosphatase stained according to the manufacturer's instructions. Briefly: the media was removed; the cells were washed with PBS; 600 µL of AP fixation solution was added to each well for 30 seconds; the cells were washed with dH₂O for 45 seconds; 600 µL of AP staining solution was added to each well and incubated at room temperature for 20 minutes in the dark; the cells were washed with dH₂O and left to dry at room temperature. The differentiation status of each colony was recorded and presented as a proportion of total colonies within that well. The differentiation status of each colony was defined as:

- <20% stained – differentiated;
- 20-90% stained – mixed;
- >90% stained – undifferentiated.

2.2.1.2.2 **Colony formation efficiency assay**

To determine the colony formation efficiency of a cell line, cells were seeded, grown and fixed as described in 1.2.1.2.1. The number of colonies/well was counted and the colony formation efficiency was calculated as follows:

$$\text{Colony formation efficiency} = \frac{\text{Number of colonies}}{\text{Number of cells seeded}}$$

2.2.1.3 *In vitro* conversion from ESCs to EpiSCs

The lab of Dr Alice Jouneau conducted the ESC-to-EpiSC conversion according to their recently published method.⁽²²⁵⁾ Briefly, prior to seeding empty 35 mm dishes were incubated with 1 mL of serum-based media for 1 hour at 37°C. The media was aspirated from the dishes, the dishes washed with PBS, and CDM media was applied (freshly supplemented with 12 µg/mL of FGF2 and 20µg/mL of Activin A). The old media was aspirated from the ESCs to be seeded, the cells washed with PBS, and trypsinised at 37°C for 2-3 minutes. An equal volume of serum-based media was applied to the trypsinised cells, and the cells dissociated by gently pipetting with a P1000 pipette. The cells were transferred to a 15 mL plastic tube and centrifuged for 5 minutes at 200 g at room temperature. The supernatant was aspirated, and the cell pellet resuspended in 1 mL of CDM. The centrifugation step was repeated and this time the cells were resuspended in the volume of media appropriate to easily plate the required number of cells based on cell counting. 1.5 million ESCs were plated into the pre-prepared 35 mm dishes coated with serum of fibronectin and containing CDM freshly supplemented with FGF2 and Activin A. The media was replaced daily.

2.2.1.4 *RT-qPCR*

2.2.1.4.1 **RNA extraction**

RNA was extracted using the RNeasy Mini kit according to manufacturer's instructions. Briefly: the cells were washed with PBS; 350 µL of RLT buffer plus 1% (v/v) 2-Mercaptoethanol was added to each well; the cells were collected in 1.5 mL tubes⁵; the cells were homogenised by passing the lysate through a 20 gauge needle and syringe; the lysate was mixed with 1 volume of 70% (v/v) ethanol and transferred to a MinElute spin column; the column was centrifuged for 1 minute at 10,000 RPM and the flow-through discarded; 350 µL of Buffer BW1 was added to the columns, the column was centrifuged for 1 minute at 10,000 RPM and the flow-through discarded; to remove contaminating DNA, 10 µL DNase I in 70 µL of Buffer RDD was added

⁵ If the lysates were not immediately extracted they were stored at -80°C.

and the columns were incubated at room temperature for 20 minutes; 350 μ L of Buffer BW1 was added to the columns, the column was centrifuged for 1 minute at 10,000 RPM and the flow-through discarded; twice, 500 μ L of Buffer RPE was added to the columns, the column was centrifuged for 1 minute at 10,000 RPM and the flow-through discarded; finally, RNA was eluted from the columns with 30 μ L of RNase-free water. The concentration of RNA was measured using a NanoDrop 2000 spectrophotometer. RNA samples were stored at -80°C .

2.2.1.4.2 cDNA synthesis

cDNA was generated using 1 μ g of RNA and the Superscript First-Strand Synthesis kit according to manufacturer's instructions. Briefly: the RNA was treated with 1 μ L of 10 mM dNTP mix, 1 μ L of oligo (dT)12-18 and made up to a final volume of 13 μ L with RNase-free water; the solution was incubated at 65°C for 5 minutes; 2 μ L of 0.1 M DTT, 4 μ L of 5X first strand buffer, and 1 μ L of RnaseOUT was added to the mixture, which was then incubated at 42°C for 1 minute; 1 μ L of 200U/ μ L Superscript II was added to each reaction and incubated at 42°C for 1 hour followed by 70°C for 15 minutes. Finally, each sample of cDNA was diluted 20 times with RNase-free and stored at -20°C .

2.2.1.4.3 RT-qPCR

RT-qPCR was conducted in either a total volume of 20 μ L with 50% (v/v) Jumpstart Sybr-Green PCR Mastermix plus 1% ROX reference dye, or in a total volume of 10 μ L with 50% (v/v) KicQstart Sybr-Green PCR Mastermix (contains ROX). RT-qPCR reactions were conducted with 300 nM of primers and 5% (v/v) of cDNA or DNA. To test for formation primer-dimer formation, a template free control was included for each primer pair on each plate. All RT-qPCR reactions were performed in triplicate on the same PCR plate. Cycling conditions were as follows: 95°C for 15 minutes; 40 cycles of 94°C for 15 seconds and 60°C for 30 seconds; 72°C for 30 seconds.

The ΔC_T and $\Delta\Delta C_T$ methods(226,227) were used to quantify the amplified target DNA. The C_T is the number of cycles required to reach a pre-defined fluorescence threshold. This threshold is set in the exponential phase of amplification. For the ΔC_T method, the geometric mean C_T

values of 2-3 housekeeping genes (*S17*, *L19* and *Gapdh*) were subtracted from the C_T value of the target gene and represented as $2^{-\Delta C_T}$. For the $\Delta\Delta C_T$ method in addition to housekeeping gene normalisation, the ΔC_T of a control sample was then subtracted, and the value represented as $2^{-\Delta\Delta C_T}$.

For ChIP-qPCR, the C_T of the target region was normalized to 1% of input DNA and to a background region and represented as fold enrichment over background.

2.2.1.4.4 **Primer design and validation**

Primers were designed using NCBI Primer-BLAST to the following specifications: amplicon size ≤ 200 bp; primer GC content 45-55%; T_m 57-63°C; and primer size 18-25 bp. RT-qPCR primers were designed to overlap exons to prevent amplification of contaminating genomic DNA. The efficiency of primer pairs was tested using 2 fold dilutions of DNA or cDNA as appropriate.(227) A good quality primer pair, suitable for analysis with the $\Delta\Delta C_T$ method, was defined as $R^2 > 0.98$ standard curve and an efficiency between 1.8 and 2.2 (Efficiency = $10^{(1/\text{slope})}$). The sequences of the primer pairs used in this thesis can be found in appendix A.1.1.1.

2.2.1.5 *ChIP-seq*

ChIP samples and sequencing libraries were prepared by Dr Rute Tomaz. Her protocols are detailed in full in her PhD thesis(228) and summarised below.

2.2.1.5.1 **ChIP**

Cells were cultured in 175cm² flasks and 1-3 flasks were harvested for each ChIP experiment. Cells were either single-fixed in 1% formaldehyde for 10 minutes at 37°C, or double-fixed by preceding the formaldehyde step by incubating in 2 mM DSG for 45 minutes. The fixation was quenched with 125 mM glycine. Cells were lysed using the buffers described in 3.1.1 and a dounce homogeniser. Nuclei were pelleted by centrifugation at 1200 RPM and resuspended in IP buffer (see 3.1.1). DNA was sonicated to 200-500 bp fragments with a Diagenode Bioruptor. The fragment size was checked by rapid de-crosslinking at 65°C and 1% (w/v)

agarose gel electrophoresis. DNA and protein samples were cleared with sepharose beads and magnetic Dynabeads by rotation for 2 hours at 4°C. DNA and protein concentration was measured with spectroscopy or Pierce BCA assay (see 3.2.1.6.2). IP was performed while rotating at 4°C overnight. Normal IgG raised in the same species as the primary antibody was used as a negative control. Chromatin was eluted from Dynabeads/sepharose beads with 250µl of elution buffer and 0.04µg/µL RNase A while shaking at 65°C overnight. Chromatin was diluted 1:2 in TE Buffer and treated with 0.2 µg/µL Proteinase K at 55°C for 2 hours. DNA was purified by phenol-chloroform extraction (see 3.2.1.5.1). DNA was either applied to qPCR (3.2.1.4.3) with primers listed in Appendix X, or used to generate sequencing libraries.

2.2.1.5.2 **Generation of libraries for sequencing**

Sequencing libraries were prepared from ChIP DNA using the NEBNext® Ultra™ DNA Library Prep Kit and Multiplex Oligos according to the manufacturer's instructions. DNA was size selected for 200 bp using two AMPure XP Beads separation steps (100 µL DNA solution and 55 µL AMPure XP Beads; 100 µL DNA solution and 25 µL AMPure XP Beads). During each size selection, the DNA solution and beads were incubated at room temperature for 10 minutes and then applied to a magnetic rack. After the second size selection, the supernatant was removed and the beads were washed 3 times with 85% (v/v) ethanol. DNA was eluted from the beads with 25 µL of 10 mM Tris-HCl pH8. 20 µL of DNA solution was PCR amplified with 1x NEBNext High Fidelity Master Mix (2X), 2.5 µL of each individual Index Primers and 2.5 µL of the Universal PCR Primer in a total volume of 50 µL. The PCR cycling conditions used were: 98°C for 30 seconds; 15 cycles of 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 30 seconds; 72°C for 5 minutes. PCR products were size selected with 40 µL of AMPure XP Beads as described above. DNA was eluted with 33 µL 10 mM Tris-HCl and the fragment size distribution was analysed in an Agilent Bioanalyzer and sequenced on an Illumina platform.

2.2.2 Computational methods

The code pertaining to the methods described here is listed in Appendix A.1.2 and can be found online at: <https://github.com/emmabell42/PhD>

2.2.2.1 *Sequence composition analysis*

The mouse genome sequence was accessed using the BSgenome.Mmusculus.UCSC.mm9 R package,(229) which contains the full mm9 mouse genome as provided by the UCSC in July 2007.

2.2.2.1.1 **GC content**

The proportion of G and C nucleotides (GC content) within a genomic region was calculated as follows:

$$GC\ content = \frac{number\ of\ C + number\ of\ G}{length - 1}$$

2.2.2.1.2 **CpG content**

The normalised fraction of CpG dinucleotides within a genomic region was calculated in accordance with the Saxonov method.(230) The method is as follows:

$$Normalised\ CpG\ fraction = \frac{Observed\ CpGs}{Expected\ CpGs}$$

The expected CpG content was calculated as follows:

$$Expected\ CpGs = \left(\frac{GC\ content}{2} \right)^2$$

2.2.2.2 *Data collection*

All publicly available microarray and sequencing datasets was downloaded from the Gene Expression Omnibus (GEO)(231). Their accession numbers are listed in appendix table A.2.

2.2.2.3 *Microarray data*

2.2.2.3.1 **Microarray processing**

The raw array intensity values were downloaded from GEO and processed in R. Each study was processed individually to minimise cross-array platform bias. Affymetrix, Illumina, and Agilent arrays were processed with the R packages *affy*(203), *lumi*(214) and *limma*(232) respectively. Briefly, the probe intensity values were $\log_2$ transformed, quantile normalised, and the Affymetrix and Agilent arrays were background corrected. Illumina arrays were only background corrected if relevant background probe intensities values were provided by the authors. Illumina BeadArray microarrays were most likely background subtracted by the Illumina BeadStudio software upon image processing, which has been found to aid detection of differentially expressed genes.(233) There is evidence that further background correction of Illumina BeadArray microarrays introduces false-positives and substantial variability, particularly at low probe intensities.(233)

To assess the statistical significance of differential gene expression, a linear regression model was fit to normalised probe intensity values from each study using the R package *limma*(232). This generated empirical-Bayes moderated t-statistics for the differential expression of a given probe between two conditions. To reconcile cross-platform probe IDs, Mouse Genome Index (MGI) gene symbols were used to identify genes. For genes with multiple probes on an array platform, the most statistically significant differentially expressed probe was used and all others discounted. In microarray experiments where a gene has been knocked out or knocked down using RNA interference (RNAi), the downregulation of that mRNA was confirmed before any further analysis.

2.2.2.3.2 **Meta-analysis of microarray data**

Meta-analyses of microarray data from ESCs upon RNAi knock down of *Esrrb* and *Nanog* were performed using the Rank Products method. Rank Products compares ranked lists of differentially expressed genes. It assumes that under the null hypothesis these lists will be ranked randomly.(234) Therefore a gene that consistently appears near the top of the list will

be considered significant due to the low probability of this occurring by chance. Rank Products meta-analysis was performed using the RankProd R package.(235) The probability of a given gene appearing in the top r of a list n genes long is calculated as $P = r/n$. The product of the rank probabilities for a given gene from each study produces an overall rank product for that gene as follows:

$$RP_g = \left(\prod_i r_{gi} \right)^{1/K},$$

where g is any given gene, $i = 1, \dots, k$, and $k = \log_2 \text{fold-change}_1 + \log_2 \text{fold-change}_2 + \dots + \log_2 \text{fold-change}_n$ (i.e. the sum of the $\log_2$ fold-changes in mRNA expression of that gene in each study that includes it) and r is the rank of the gene. The P-value and FDR for each gene is calculated by permuting the $\log_2$ fold-changes I times (here $I = 100$) and repeating $RP_g I$, to generate a reference distribution.

2.2.2.3.3 Gene enrichment analysis of microarray data

Gene enrichment analysis was performed as described in 2.2.2.5.2 on the t-statistics from each microarray included in the meta-analyses. The resulting False Discovery Rate (FDR) corrected P-values were combined using Fisher's method(236) to generate an overall statistical evaluation of the differential expression of each pathway. Fisher's method of combining P-values was performed by calculating the following summary statistic:

$$S = -2 \sum_{k=1}^k \log(P_{ik}),$$

where k is the number of studies that included that gene and P_i is the P-value of that gene in a given study. A P-value for the summary statistic was derived from the χ^2 distribution with $2k$ degrees of freedom.

2.2.2.4 Sequencing data

2.2.2.4.1 Sequencing data processing

2.2.2.4.1.1 Sequencing read processing

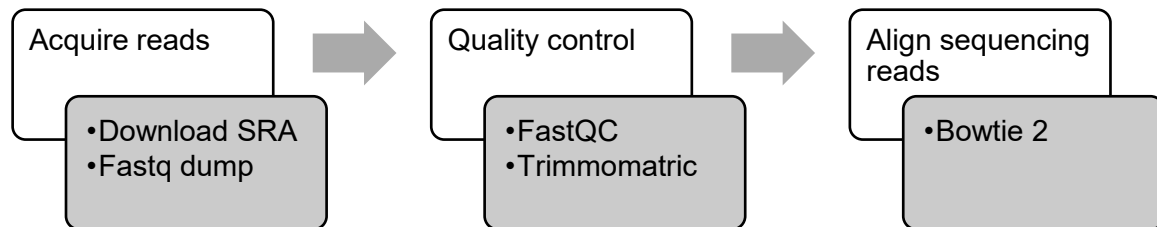


Figure 2.1 An overview of the sequencing read processing pipeline.

Sequencing reads were downloaded from GEO as Sequence Read Archives (SRAs) and converted to fastq format using Fastq Dump.(237) Read quality was assessed using FastQC.(207) Read filtering and trimming of poor quality bases and any adaptor or otherwise contaminating sequences identified by FastQC was performed with Trimmomatic using the default settings.(223) Reads were aligned to the mm9 mouse reference genome(43) using Bowtie2.(205,206) Processed sequencing data was visualised as bed and BigWig(238) files in the UCSC Genome Browser (figure 2.1).(239)

2.2.2.4.1.2 Bisulphite-seq

Bisulphite-sequencing and data processing was performed by Drs Alice and Luc Jouneau as described in Habibi *et al*, 2013.(133) CpGs with coverage of less than 7 reads were filtered from the dataset. CpG methylation status (i.e. unmethylated or methylated) was determined using a Hidden Markov Model (HMM) algorithm using the TileHMM R package(240).

2.2.2.4.1.3 ChIP-seq

Statistically significantly enriched peaks were identified using MACS2(215,241) by comparing the read densities of the ChIP-seq experiment to the relevant control ChIP-seq sample. These controls were typically matched ChIP-seq samples that were not treated with an antibody (input controls), however a few experiments used an IgG or GFP antibody. Only datasets

including a control experiment were used in this study. When running MACS2, shift-size modelling was enabled. If the shift-size was estimated at <70, peak calling was repeated without modelling, and the shift-size was set based on the fragment size selection stated in the publication associated with the ChIP-seq experiment. For histone modification ChIP-seq experiments, MACS2 was run with the “--broad” parameter, which creates composite peaks from neighbouring read dense regions.

2.2.2.4.1.4 **DIP-seq**

Methylated (MeDIP), hydroxymethylated (hMeDIP) and formylcytosine (fMeDIP) DNA Immunoprecipitation sequencing experiments, collectively termed here DIP-seq, were analysed using the *MEDIPS* R package.(242) Processed sequencing reads were read into R as MEDIPS SET objects using mm9 as the reference genome (parameters: BSgenome="BSgenome.Mmusculus.UCSC.mm9", uniq=1e-3, extend=300, ws=100). One of these MEDIPS SET objects was selected at random to create a coupling set object, counting the number of CpGs within each genomic window. The differential coverage between a DIP-seq experiment and its input control was calculated using the *MEDIPS.meth* function (parameters: p.adj="bonferroni", MeDIP=T). This gives a Relative Methylation Score (RMS) for each window, normalised for CpG density.

2.2.2.4.1.5 **RNA-seq**

RNA-seq was performed by Dr Alice Jouneau's group.(130) Briefly, cDNA was quantified by Qubit (Invitrogen), 20 ng of cDNA was size selected (~300 bp) and amplified by PCR (14 cycles). The resulting DNA product was quality checked by Bioanalyzer (BioRad). Sequencing was performed using the Illumina HiSeq according to the standard Illumina protocol. Data processing was performed using the Illumina Analysis Pipeline (1 mismatch, uniquely aligned reads only). The outputted BED files were normalised for sequencing depth and mapping efficiency by scaling the total number of sequencing reads per file to the sample with the smallest number of reads.

2.2.2.4.2 Sequencing data analysis

2.2.2.4.2.1 Gene annotation and motif enrichment

Gene annotation and motif enrichment was performed using HOMER, using their standard protocols.(212)

2.2.2.4.2.2 Read coverage comparison

Sequencing read comparison was performed in R using the custom *plotCoverage* function written by Dr Edward Curry. This function is given in appendix A.2.1. BAM files were subset to the regions of interest using Samtools.(221) The subsetted BAM files were read into R using the *GenomicAlignments* package. Duplicate reads were identified based on duplicate read start position and filtered. BAM files were normalised for sequencing depth as follows:

$$C_1 = C_0 \times \frac{10^6}{N}$$

Where C_0 is the read coverage object, N is the number of reads in C_0 , and C_1 is the sequencing depth normalised read coverage object.

Sequencing depth normalised read coverage objects were then normalised to their control ChIP-seq. A pseudocount of 1 was added to each sequencing depth read normalised coverage object to avoid division by zero.

$$C_2 = \frac{\text{Experiment } C_1 + 1}{\text{Control } C_1 + 1}$$

Where C_1 is the sequencing depth normalised read coverage object, and C_2 is the control normalised read coverage object.

Finally to calculate and visualise the read coverage at 2 or more sets of regions, the *plotCoverage* function was used. Each set of regions was divided into bins and the average number of reads within each bin was calculated. For the final visualisation, 10,000 bins were used.

2.2.2.5 General analyses

2.2.2.5.1 Gene ontology

Gene ontology was performed with PANTHER DB(243) PANTHER Overrepresentation Test, using the PANTHER GO-Slim or GO-Complete Biological Process annotation dataset, compared to the *Mus musculus* reference gene set. P-values were Bonferroni corrected for multiple hypothesis testing.

2.2.2.5.2 Gene enrichment analysis

Gene enrichment analysis was performed using the *Limma* R package.(232) To test whether the genes in a given pathway tend to be differentially expressed compared to a random set of genes, lists of mouse pathways and their constituent genes were sourced from the Consensus Path Database (CPDB).(244,245)

2.2.2.5.3 Clustering analysis

At several points during this study, hierarchical clustering was used to mathematically evaluate the relationship between different experiments. Typically, a distance metric of 1 minus the Pearson correlation coefficient between sets of statistics (e.g. gene-wise profiles of differential expression t-statistics) was used to capture similarities/differences between the different experiments. A matrix of pair-wise distance scores was constructed, and visualised using a heatmap with hierarchically clustered rows and columns.

Principal Components Analysis (PCA) was also used to assess experimental variation. An input table (e.g. gene-wise profiles of differential expression t-statistics) was quantile normalised and any missing t-statistics were imputed by *k*-nearest neighbour averaging, where *k*=10, using the R package *impute*. (246)

3 RESULTS I

DNA METHYLATION PROFILING OF IN VITRO PLURIPOTENCY MATURATION

3.1 Introduction

The ESC transcriptional network centres on the OSN triad.(20,24–26,29) These proteins auto- and cross-regulate to promote their own transcription, as well as that of other self-renewal promoters.(247) Simultaneously, they inhibit the expression of differentiation drivers. Perturbing the ESC transcriptional network causes its collapse,(17,18) triggering specific or mixed lineage differentiation.(20,248) Furthermore, incubating MEFs with OCT4 and SOX2, alongside KLF4 and MYC reprograms them to pluripotent cells similar to ESCs, termed Induced Pluripotent Stem Cells (iPSCs).(170) In summary, the ESC transcriptional network both maintains and induces pluripotency. In terms of their mechanics: these transcription factors form multi-protein complexes at gene promoters and enhancers; (17,18) interact with chromatin modifiers such as the Mediator and Cohesin complexes, and P300;(180,249) and they recruit the transcriptional machinery to gene promoters to trigger transcription.(249)

There exists a second source of pluripotent stem cells in the developing mouse embryo: EpiSCs.(10) While both pluripotent, ESCs and EpiSCs display distinct properties and behaviours. ESCs and EpiSCs represent a more naïve state of pluripotency and one primed for differentiation respectively.(6) ESCs and EpiSCs maintain distinct yet overlapping transcriptomes. Both cell types express *Pou5f1* (OCT4), *Sox2* and *Nanog*.(10,11) However, key naïve pluripotency genes lose expression in EpiSCs (e.g. *Esrrb*; *Dppa3*, *Klf4*, *Tbx3* and *Zfp42*(10)), while primed pluripotency genes gain expression (e.g. *Fgf5* and *Nodal*(11); *Otx2*(10)). In contrast to ESCs, we know relatively little about the EpiSC transcriptional network, and how they maintain their primed pluripotency.

As well as regulating the genome at the DNA sequence level, mammalian cells control gene expression via epigenetic mechanisms that modify nucleotides and the histone proteins that package DNA. CpG methylation of promoters(67–73) and enhancers(74–80) correlates with chromatin condensation and transcriptional repression. Adding methyl and acetyl moieties to specific amino acids within histone proteins condenses and decondenses chromatin.(250) For example, H3K27ac associates with an open chromatin state,(251,252) whereas H3K27me3 associates with a closed state.(252,253) Re-organisation of the epigenome accompanies the shift in transcriptome and cell signalling pathway dependency during pluripotency maturation.(5,10,11,129,130) For instance, CpG methylation displays a rapid global increase in pre-implantation blastocysts,(123,124) predominantly in the ICM and not the TE.(124) Thus, one can use both transcriptional and epigenetic markers to assay gene activity and pluripotency maturation.

Super Enhancers drive high expression of cell-specific genes.(179,194,195) Super Enhancers, in contrast with mere 'Typical Enhancers', contain the histone modification H3K27ac – and in ESCs, the Mediator Complex component MED1 – at unusually high density.(179) ESC Super Enhancers and Typical Enhancers bind similar densities of OCT4, SOX2 and NANOG, however ESRRB and KLF4 exhibit preferential enrichment at Super Enhancers.(179) Whyte *et al.*, found Super Enhancers drive the expression of tissue-specific genes.(179) In ESCs many of the targets of Super Enhancers comprise genes previously identified as self-renewal and pluripotency regulators.(179) Whether these enhancers remain active or are decommissioned in EpiSCs and during the transition between naïve and primed pluripotency remains unclear.

In this chapter, I aim to determine whether Super Enhancers regulate the primed EpiSC transcriptome or only represent naïve ESC pluripotency. If so one may use them to model enhancer decommissioning during pluripotency maturation. I replicate Whyte *et al.*'s, observations using publicly available sequencing and microarray datasets, as well as newly generated bisulphite sequencing datasets generously shared by our collaborators Dr Alice

Jouneau and Prof Henk Stunnenberg. I develop my own sequencing and microarray data processing pipelines in order to do so. I replicate Whyte *et al*'s., finding that naïve pluripotency associated transcription factors (e.g. ESRRB and KLF4) bind at higher density to Super Enhancers compared to Typical Enhancers. I then use these datasets to further characterise the epigenetic regulation of ESC Super Enhancers and Typical Enhancers in ESCs and EpiSCs. Specifically, I use CpG methylation – a key epigenetic remodelling event during this developmental transition – to identify active and inactive enhancers. Finally, I determine whether ESC Super Enhancers continue to drive transcription in EpiSCs, by analysing their epigenetic profiles and the transcription of their associated genes.

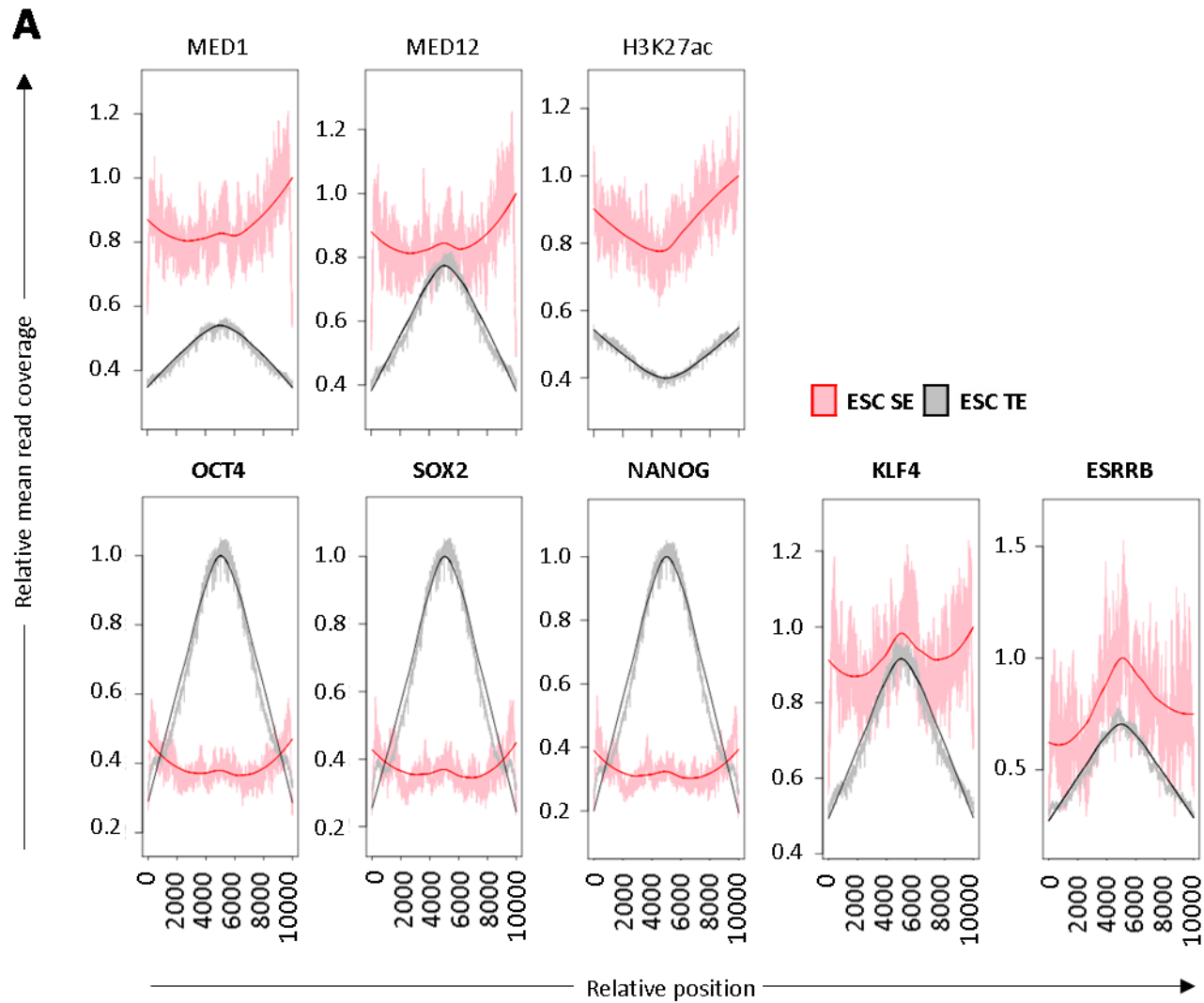
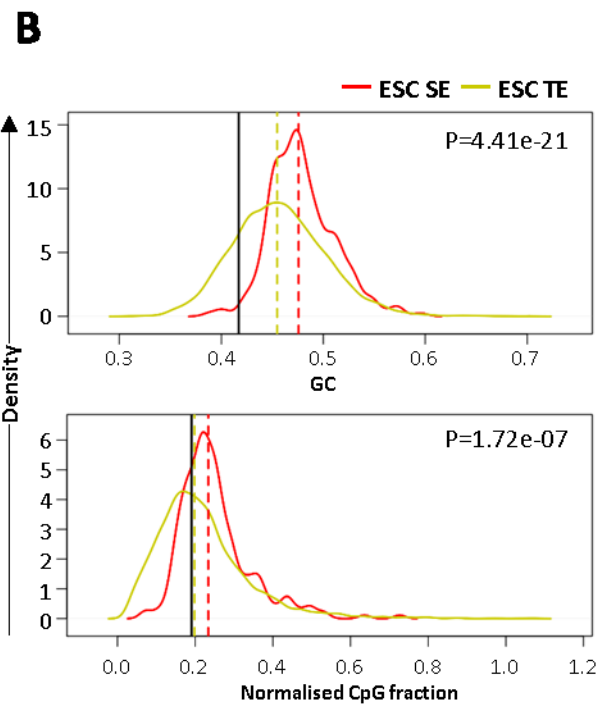
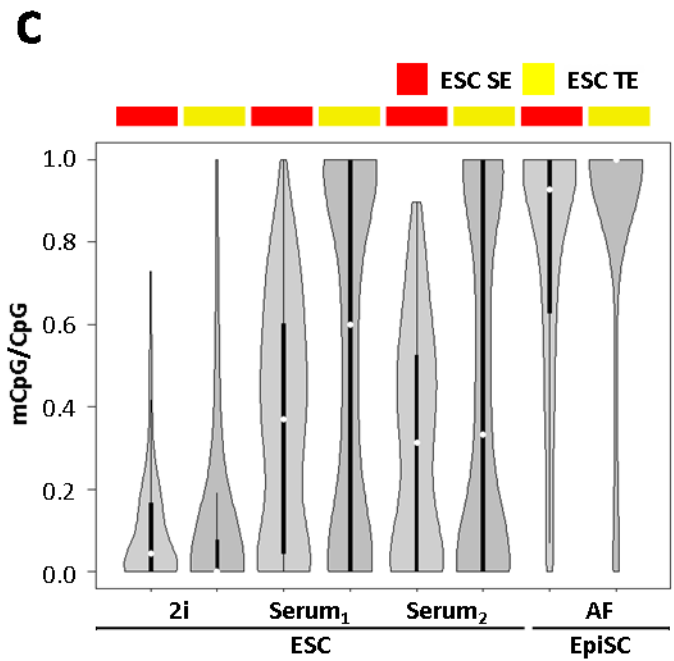
3.2 Results

3.2.1 ESC Super Enhancers and Typical Enhancers appear CpG methylated in EpiSCs

To further analyse transcription factor binding at the ESC Super Enhancers published by Whyte *et al.*'s,(179), I developed a pipeline to plot metagenes of sequencing read coverage at a given set of genomic regions.⁶ This pipeline downloads, trims and filters, aligns, normalises for sequencing depth and, if relevant, background signal, and finally outputs a 'metagene' showing relative average read coverage at given sets of genomic regions. Dr Edward Curry wrote the *plotCoverage* R function used in the last step (appendix A.2.1). To control for the differing lengths of enhancer regions, the *plotCoverage* function calculates the average sequencing read coverage over 10,000 bins across the length of the genomic regions of interest. Figure 3.1A shows a panel of metagenes comparing ChIP-seq read coverage at ESC Super Enhancers and Typical Enhancers. For each panel the maximal sequencing read coverage was designated as an arbitrary unit of 1 and the mean sequencing read coverage is plotted relative to that point. The MED1 plot recapitulates the MED1 coverage observed by Whyte *et al.*, at ESC Super Enhancers and Typical enhancers, evidencing that my pipeline and Dr Curry's *plotCoverage* function both work correctly.

As Whyte *et al.*, found, in addition to MED1, the ESC pluripotency factors KLF4 and ESRRB bind more densely to Super Enhancers than Typical enhancers. Whyte *et al.*'s., method of identifying Super Enhancers requires knowledge of the master transcription factors in the cell type of interest. We do not yet possess this knowledge for many cell types. In order to bypass this prerequisite, Hnisz *et al.*, searched for a surrogate enhancer marker. They found that using H3K27ac reproduced most ESC Super Enhancers, while minimising excess regions.(195)(195) Similarly to MED1, I found that H3K27ac and the Mediator complex

⁶ Methods: Chapters 2.2.2.4.1 and 2.2.2.4.2.2



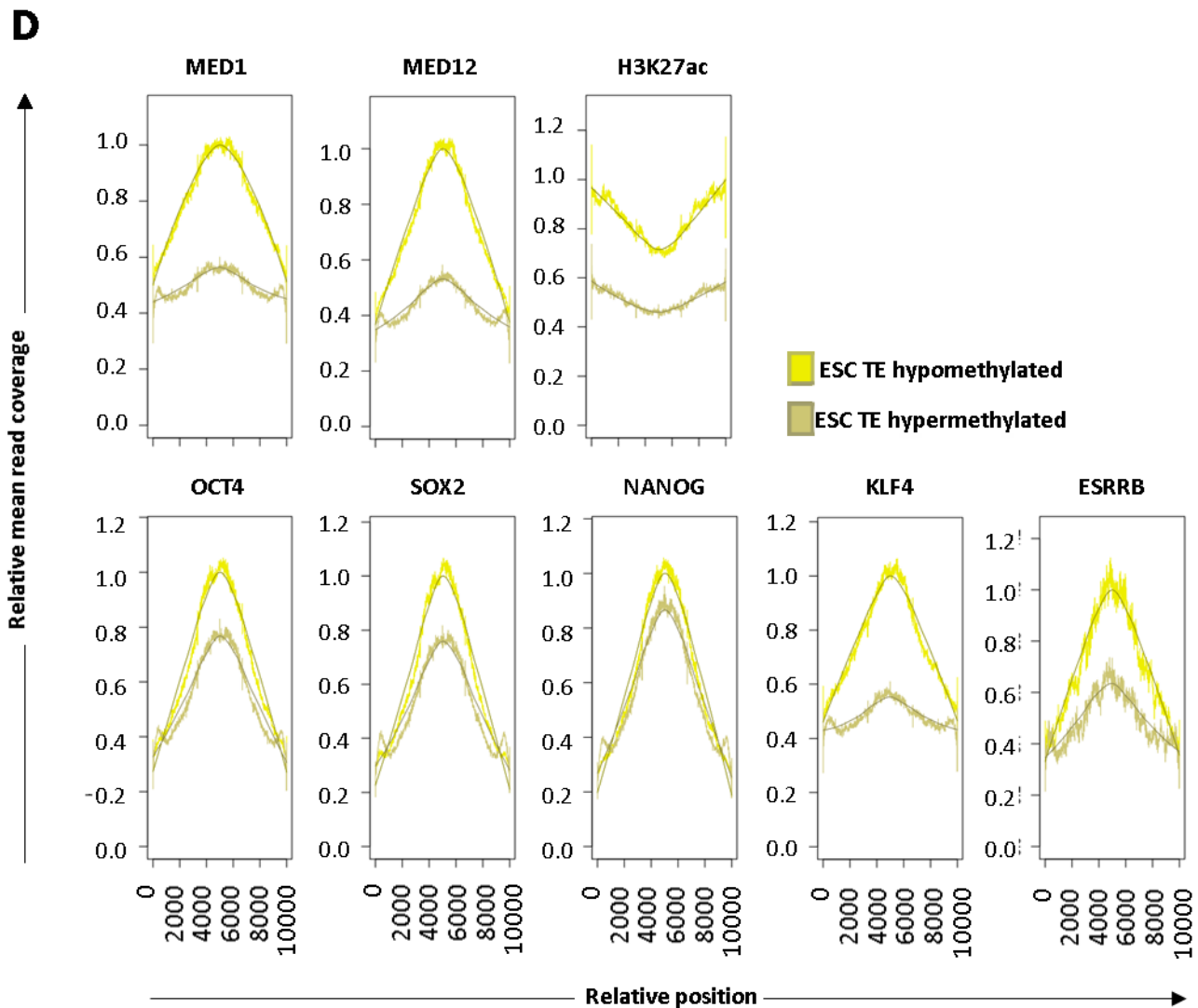


Figure 3.1 ESC Super Enhancers and Typical Enhancers acquire CpG methylation as pluripotency matures.

A. Metagenes showing the mean ChIP-seq read coverage at ESC Super Enhancers (SE) and Typical Enhancers (TE). ChIP-seq bam files were normalised for sequencing depth and background. The mean read coverage values at ESC Super Enhancers and Typical Enhancers was calculated using the `plotCoverage` function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value for either region set. The darker lines show the smoothed fit of the data generated by loess regression. **B.** Density plots comparing GC content and normalised CpG fraction at ESC Super Enhancers and Typical Enhancers. The dotted lines indicate the medians for each region set and the black solid lines indicate the global average for the mouse genome. **C.** Violin plots showing the mCpG/CpG at ESC Super Enhancers and Typical Enhancers during pluripotency maturation. The ESC-2i and ESC-serum1 data come from Habibi et al.,(3) the ESC-serum2 data comes from Stadler et al.,(4), and the EpiSC data comes from Dr Alice Jouneau. **D.** Metagenes showing the mean ChIP-seq read coverage at hypomethylated and hypermethylated ESC Typical Enhancers (TE). These were generated as detailed in legend 3.1.A.

member MED12 occur at higher density at ESC Super Enhancers relative to Typical Enhancers. In contrast to MED1, H3K27ac, ESRRB and KLF4, OSN appears to bind at higher density to ESC Typical Enhancers compared to Super Enhancers. The composition of Super Enhancers and Typical Enhancers may explain this. Typical Enhancers consist of a single constituent enhancer, whereas Super Enhancers contain multiple constituent enhancers. The single focal point of Typical Enhancers compared to the multiple scattered focal points of Super Enhancers, may cause the appearance of mean higher binding density at Typical Enhancers. If a transcription factor binds at even density across a Typical Enhancer and the distinct constituents of a Super Enhancer, then the average binding density over the whole Super Enhancer will appear lower due to inclusion of interstitial non-enhancer regions with lower binding density.

CpG methylation strongly negatively correlates with enhancer activity. Furthermore, Wiench *et al.*, found CpG methylation patterns could define cell-specific and constitutively active enhancers.⁽⁷⁷⁾ To test whether CpG methylation might provide a line of evidence as to Super Enhancer activity, I first compared the GC and CpG content of ESC Super Enhancers and Typical Enhancers.⁷ I calculated the CpG content as the normalised CpG fraction of observed/expected CpGs. Whyte *et al.*, observed that compared to Typical Enhancers, Super Enhancers occupy genomic regions an order of magnitude longer. I therefore normalised both the GC and CpG content to the length of the enhancer. Overall, Super Enhancers contain both a greater proportion of GC bases (figure 3.1B upper panel) and normalised CpG fraction (figure 3.1B lower panel). In fact, the median normalised CpG fraction in Typical Enhancers approximates the global average (black line). However, ESC Typical Enhancers show greater range in both GC and CpG content.

During early embryonic development pluripotent stem cells reorganise their epigenome.^(10,123,124) Their difference in CpG content implies that CpGs within ESC Super Enhancers compared to Typical Enhancers may experience stronger evolutionary

⁷ Methods: Chapter 2.2.2.1

selection⁽⁵⁶⁾ and/or hypomethylation during early embryonic development.⁽⁵⁵⁾ Therefore, I posited that ESC Super Enhancers and Typical Enhancers might display differing CpG methylation profiles during pluripotency maturation. To test this hypothesis I used publicly available bisulphite-seq data from ESCs cultured under naïve (2i+LIF) and priming (serum+LIF) conditions, as well as a newly generated EpiSC dataset (figure 3.1C). To calculate the proportion of methylated CpGs within each enhancer, Drs Alice and Luc Jouneau used a Hidden Markov Model (HMM) to assign each CpG as either unmethylated or methylated.⁸ In agreement with the literature, under 2i+LIF conditions both enhancer classes displayed minimal CpG methylation.⁽¹³³⁾ In ESCs grown in serum+LIF, both sets of enhancers acquire CpG methylation, however they do so in markedly different patterns. Typical Enhancers display a stark bimodal distribution of mCpG/CpG with approximately even numbers of enhancers displaying <0.25 and >0.75 mCpG/CpG. In combination with the lower GC content and normalised CpG fraction, suggests that the hypermethylated subgroup of Typical Enhancers may not represent active enhancers. Moreover, the hypermethylated subgroup show lower density of the enhancer defining Mediator complex (MED1 and MED12), and H3K27ac, as well as the ESC pluripotency associated transcription factors KLF4 and ESRRB (figure 3.1D). Whyte *et al.*, defined Typical Enhancers as low density MED1 binding regions. This method may capture inactive regions to which MED1 sporadically binds and/or false-positive peak calling events. Thus the low MED1 binding further substantiates the argument that hypermethylated Typical Enhancers do not represent active enhancers in ESCs. Super Enhancers also acquire CpG methylation in a bimodal distribution, however the hypermethylated group of enhancers falls between $0.25 < \text{mCpG/CpG} < 0.8$. Finally, in EpiSCs, both Super Enhancers and Typical Enhancers display marked hypermethylation. This may indicate that both sets of enhancers no longer drive transcription.

⁸ Methods: 2.2.2.4.1.2

3.2.2 Dormant Super Enhancers appear hypermethylated in ESCs

CpG methylation increases when ESCs differentiate to EpiSCs (135,136) due to increased expression of the *de novo* methyltransferase DNMT3B.(137) Therefore, the increase in CpG methylation at ESC Super Enhancers may simply reflect this transition. To discount this we decided to compare the CpG methylation profiles of ESC Super Enhancers and Super Enhancers from a murine somatic cell type in ESCs and EpiSCs. One would presume that Super Enhancers defined in a somatic cell type would not drive gene expression until activated later on during development. In addition to ESCs, Whyte *et al.*, defined Super Enhancers in myoblast cells, macrophages, and lymphocytes (ProB and Th cells).(179) To decide which set of Super Enhancers would provide the most appropriate comparison we assessed their lengths, GC content and normalised CpG fraction (figure 3.2A and 3.2B). Myoblast, macrophage and ProB cell Super Enhancers appear the most similar in length. Myoblast, ProB cell and Th cell Super Enhancers appear the most similar in GC content. ProB and Th cell Super Enhancers appear the most similar in normalised CpG fraction. As ProB cell Super Enhancers appear the most similar to ESC Super Enhancers overall, we selected these as our candidate murine somatic Super Enhancers for comparison. We calculated the mCpG/CpG upstream, downstream, and at the Super Enhancers themselves (figure 3.2C), defining the length of the upstream and downstream regions as 10% of the length of each Super Enhancer. At ESC Super Enhancers, the enhancer displays hypomethylation relative to the flanking regions. In contrast, the ProB cell Super Enhancers show equally high CpG methylation at all three regions. This suggests ESCs actively maintain a relatively unmethylated state at ESC Super Enhancers. The progressive CpG methylation of ESC Super Enhancers may evidence their activity in ESCs and decommissioning in EpiSCs.

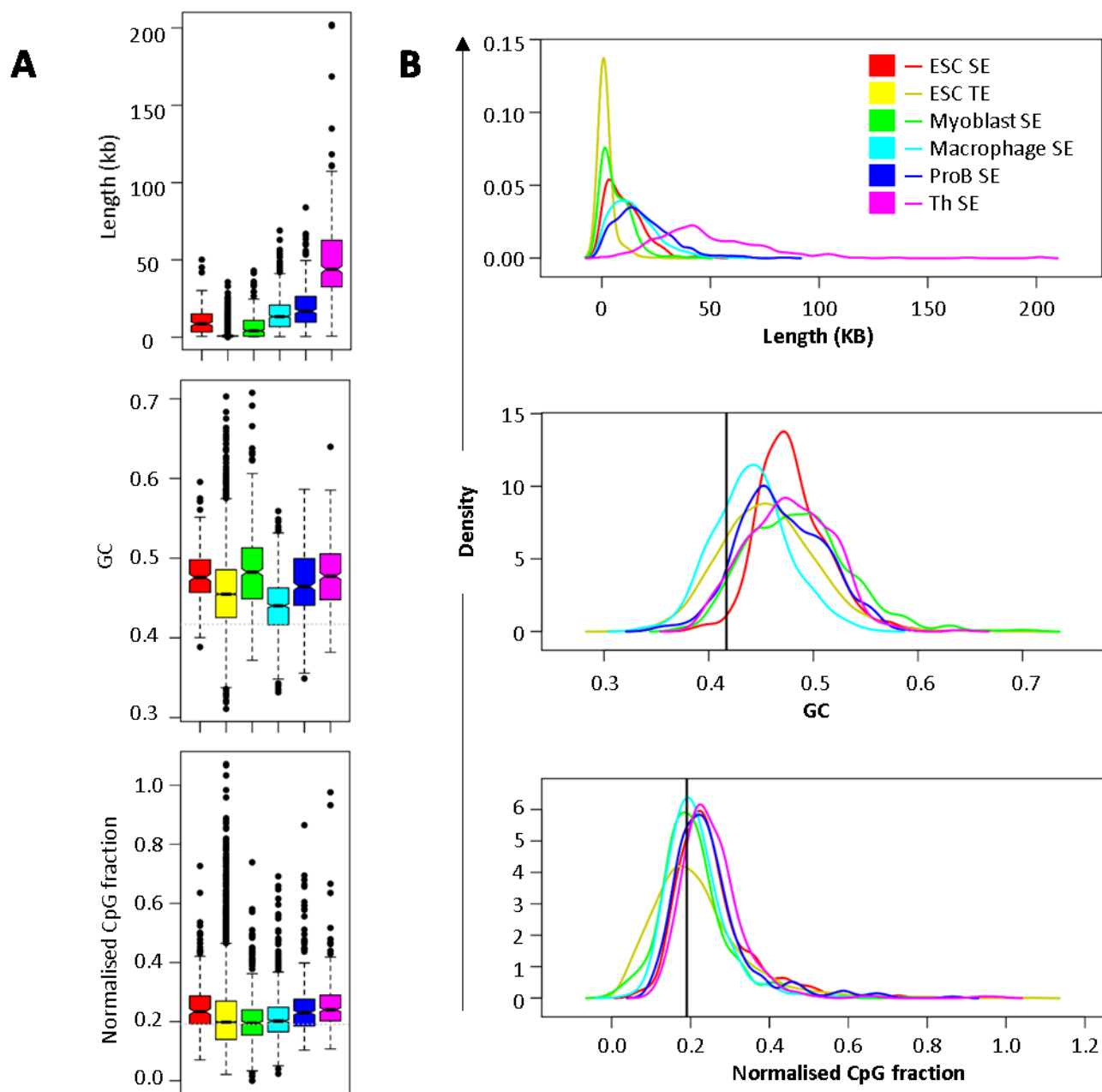
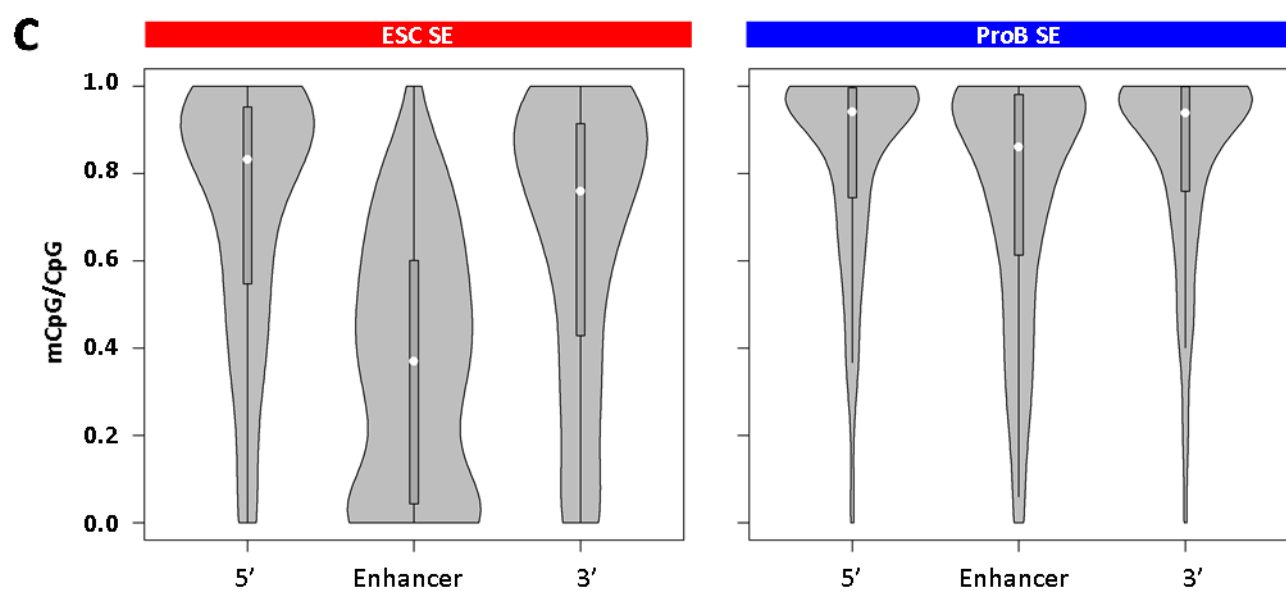


Figure 3.2 ESC but not ProB Super Enhancers appear selectively CpG hypomethylated in ESCs.

The key in section B applies to all panels. **A.** Boxplots comparing the length, GC and CpG content of the enhancers defined by Whyte *et al.* **B.** Density plots comparing the length GC and normalised CpG fraction of the enhancers defined by Whyte *et al.* The black line on the second and third plots indicates the global average for the mouse genome. **C.** CpG methylation at Super Enhancers and their 5' and 3' flanking regions. Flanking regions were designated as 10% of the length of each Super Enhancer upstream/downstream of the start/end co-ordinate.

3.2.3 Transcription factor binding at ESC Super Enhancers clusters at hypomethylated constituent enhancers

Whyte *et al.*'s, definition of Super Enhancers joins together enhancers within a 12 KB window.⁽¹⁷⁹⁾ Therefore, most Super Enhancers comprise multiple constituent enhancers bordered by interstitial regions. A closer look at the *Klf4*-associated Super Enhancer reveals that in ESCs grown in serum +LIF, CpG methylation sharply demarcates these two subregions (figure 3.3A). The CpGs within the constituent enhancers form an unmethylated block bordered by the methylated interstitial regions. To validate this pattern of CpG methylation, I used publicly available DNA Immunoprecipitation (DIP-seq) sequencing targeting 5-mC (figure 3.3B).⁽¹²⁹⁾ I processed the DIP-seq data with the R package MEDIPS⁹.⁽²⁴²⁾ MEDIPS corrects for CpG density bias and generates a Relative Methylation Score (RMS) between 0-1 summarising the methylation status of all CpGs within each 100 bp window within the genome. Thus, the RMS reflects both heterogeneity amongst the CpGs in each 100 bp window, and that of the cell population (i.e. in culture some cells may methylate a given CpG, while other may not). In the ESC lines E14 and J1, the subregions show a similar pattern of CpG methylation: the 'unmethylated' subregions appear largely hypomethylated relative to the 'methylated' subregions. The 'unmethylated' subregions also appear more highly methylated in EpiSCs compared to ESCs (Figure 3.3B, upper panel). In parallel, I examined 5-mC methylation profiles in extra-embryonic stem cells derived from the trophecterm (TS; trophoblast stem cells) and the primitive endoderm (XEN; extra-embryonic endoderm stem

⁹ Methods: Chapter 2.2.2.4.1.4

cells) of the early embryo (Figure 3.3B). Both subregions appear hypermethylated in TS and XEN cells, indicating that ESCs uniquely maintain these constituent enhancers hypomethylated at this early developmental stage.

The TET family of enzymes oxidise 5-mC to 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), and finally 5-carboxylcytosine (5-caC).^(60,63) Mouse and human ESCs abundantly express TET1 and TET2,⁽²⁵⁴⁾ and unlike most cell types contain high concentrations of 5-hmC.^(60,255) As pluripotency matures 5-hmC rapidly declines, while 5-mC increases.^(256,257) In contrast to 5-mC, in ESCs 5-hmC shows enrichment in gene-rich euchromatin (e.g. promoters, enhancers, and gene bodies).^(255,256,258) This enrichment is associated with 'poised' chromatin marked by the bivalent chromatin signature H3K27me3 and H3K4me3,^(127,128) as well as active enhancers.^(259,260) TET1 may specifically induce 5-hmC at enhancers.⁽²⁶¹⁾ Bisulphite treatment of genomic DNA converts unmethylated cytosine nucleotides to uracil. Both 5-mC and 5-hmC bases escape this conversion. Bisulphite-seq therefore cannot differentiate between 5-mC and 5-hmC. To determine the type of methylation at 'methylated' ESC Super Enhancer subregions, I analysed two DIP-seq experiments targeting 5-mC, 5-hmC, and 5-fC in ESC-serum (figure 3.3C). As shown in figure 3.3B, compared to 'unmethylated' subregions, 'methylated' subregions harbour higher levels of 5-mC. This trend appears more pronounced for 5-hmC and to some extent 5-fC at the same 'methylated' subregions. Together, this suggests an active de-methylation process may operate across Super Enhancers to maintain an intermediate CpG methylation level as compared to their flanking regions (see figure 3.2C).

In addition to MED1, many other transcription factors more densely occupy the constituent enhancers within the *Klf4*-associated Super Enhancer compared to the interstitial regions (figure 3.3A). To test whether this affects all ESC Super Enhancers, I generated metagenes plotting the relative mean sequencing read coverage of a panel of key ESC pluripotency associated transcription factors including ESRRB, KLF4, OCT4, SOX2, and NANOG at

unmethylated and methylated Super Enhancer subregions and ESC Typical Enhancers (figure 3.3D). Each of these molecules displays higher normalised sequencing read coverage at

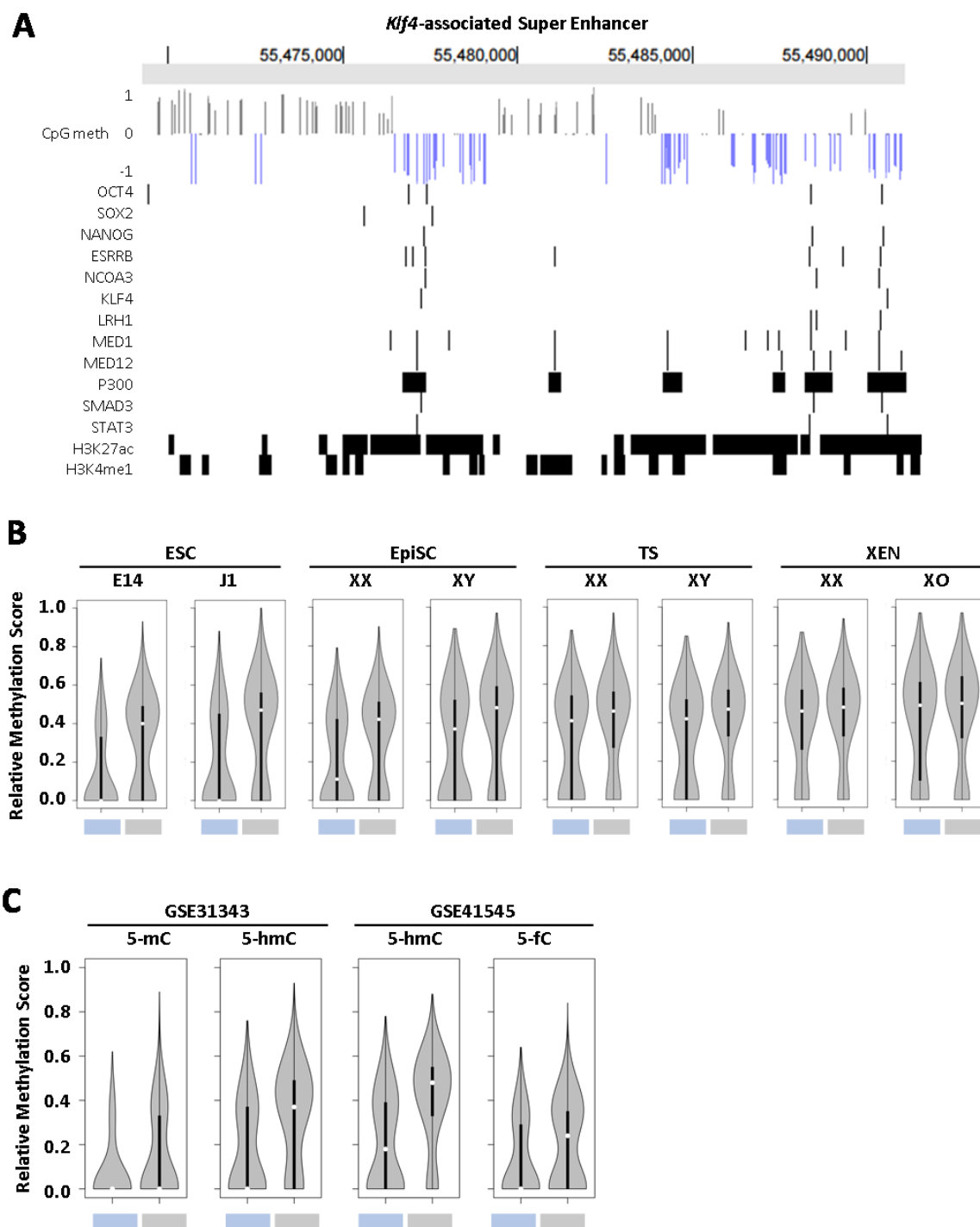


Figure 3.3 CpG hypomethylation delineates the constituent enhancers within ESC Super Enhancers.

Full legend on page 61.

D

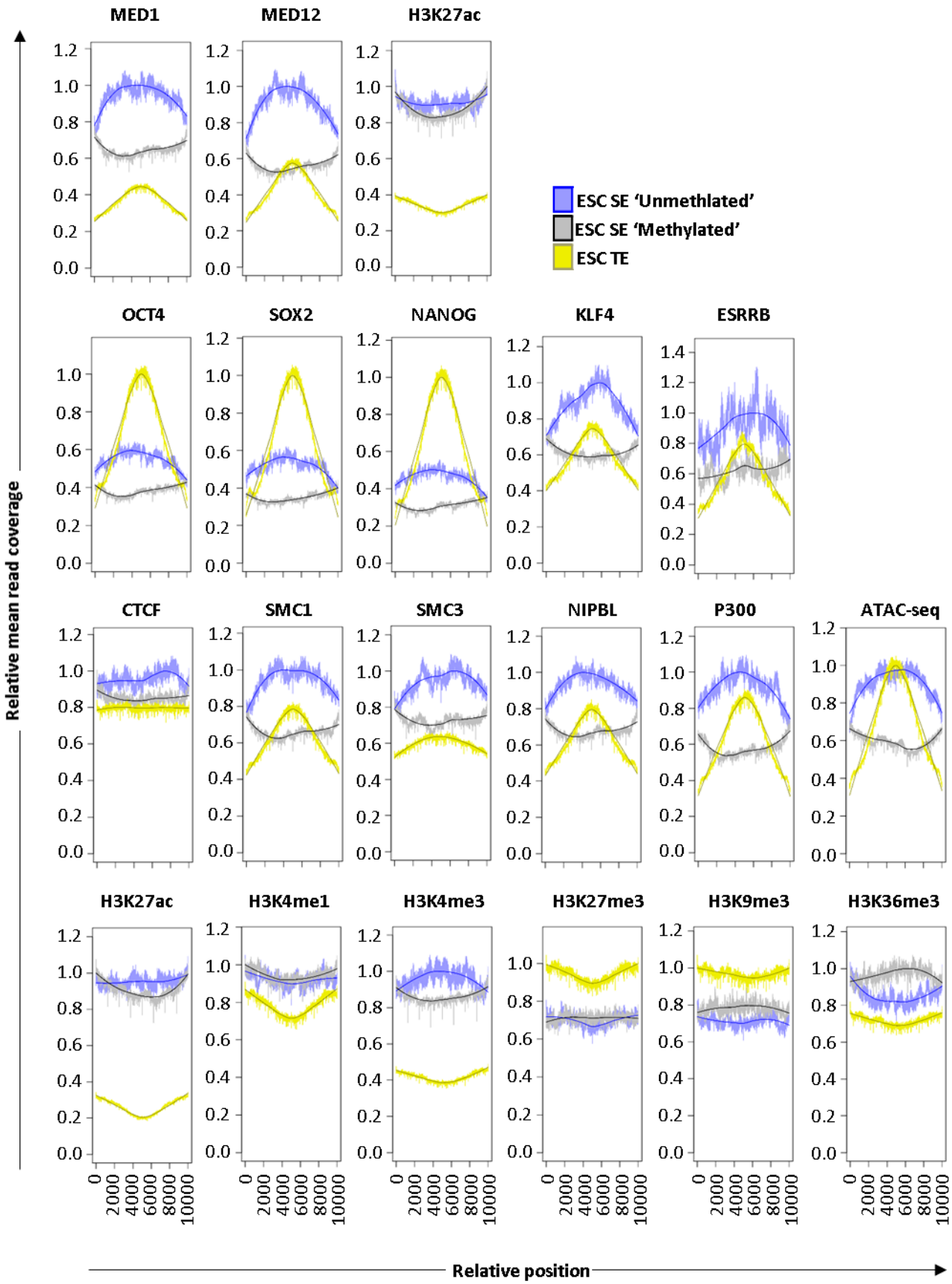


Figure 3.3 CpG hypomethylation delineates the constituent enhancers within ESC Super Enhancers.

A. Screenshot of the *Klf4*-associated Super Enhancer showing CpG methylation as well as ChIP-seq peaks for key pluripotency transcription factors and enhancer associated histone modifications. CpG methylation (CpG meth) is presented on a -1-1 scale. For each CpG, the methylation state (i.e. unmethylated or methylated) was called using a Hidden Markov Model (HMM). The colour and direction of each bar indicates the HMM assessed methylation status (unmethylated = down, blue, methylated = up, grey). The height of the bar indicates the β -value, which is a statistically evaluated methylation score ranging from 0-1. Here, the β -values pertaining to unmethylated CpGs have been subtracted from 1 and multiplied by -1 to convert them to a -1-0 scale. The ChIP-seq data visualised here is listed in appendix table A.3. **B.** DIP-seq for 5-mC in ESCs, EpiSCs, TS, and XEN cells from Senner et al., at unmethylated and methylated ESC Super Enhancer subregions. CpG methylation is represented as a MEDIPS derived Relative Methylation Score, which corrects for CpG density bias. **C.** DIP-seq for 5-mC, 5-hmC, and 5-fC in ESCs at unmethylated and methylated ESC Super Enhancer subregions. CpG methylated is represented as detailed in figure legend 3.3.B. **D.** Metagenes showing the mean ChIP-seq read coverage at Unmethylated and Methylated ESC Super Enhancer subregions and ESC Typical Enhancers. ChIP-seq bam files were normalised for sequencing depth and background. The mean read coverage values at ESC Super Enhancers and Typical Enhancers was calculated using the `plotCoverage` function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value for either region set. The darker lines show the smoothed fit of the data generated by loess regression.

unmethylated compared to methylated subregions, along with essential enhancer complexes such as Mediator (MED1 and MED12) (figure 3.3D, first row), Cohesin (SMC1, SMC3 and NIPBL) and the H3K27 histone acetylase P300 (figure 3.3D, third row). To compare chromatin accessibility between the two subregions, I plotted relative mean sequencing read coverage of an ATAC-seq experiment. The ATAC-seq metagene replicates the pattern observed in the transcription factor metagenes, indicating that compared to the methylated subregions, the unmethylated subregions maintain more accessible chromatin (figure 3.3C, third row). The clustering of transcription factors(17,18) and the Mediator and Cohesin complexes(180,249) to the unmethylated subregions, alongside their increased chromatin accessibility, further evidences unmethylated subregions act as enhancers.

3.2.4 ESC Super Enhancers physically interact with their proposed associated genes in ESCs

Enhancers drive gene expression, therefore to meaningfully assess the activity of an enhancer one must consider its associated genes. Whyte *et al.*, associated Super Enhancers with genes by determining the most highly expressed gene within 50kb. They validated selected interactions by comparison to previously generated 3C-qPCR and HiC-seq data. Using their own promoter capture HiC-seq, our collaborators Dr Henk Stunnenberg and Dr Wout Megchelenbrink re-assessed the validity of these proposed associated genes. Of the 231 ESC Super Enhancers, 56 show no detectable promoter interaction, 54 interact with a single promoter, and 121 interact with multiple promoters. For those Super Enhancers that detectably interact with a promoter, 114 of 175 interact with the promoter assigned by Whyte *et al.*, which equates to 65% (figure 3.4A). It is unclear whether the undetected Super Enhancer-promoter interactions result from the limitations of the technology, experimental set-up or simply do not occur. One explanation for non-detected interactions stems from the observation that many ESC Super Enhancers overlap with gene bodies (figure 3.4B). In some cases (e.g. *Esrrb*, *Tet2*), they overlap with the gene Whyte *et al.*, suggested they regulate. Highly proximal Super Enhancer and promoters would prove difficult to detect by HiC-seq due to resolution limitations. We concluded that Whyte *et al.*'s, list sufficiently reflects the genes that ESC Super Enhancers regulate to detect broad changes in transcription due to their activity.

3.2.5 ESC Super Enhancer associated gene promoters acquire CpG methylation during pluripotency maturation

Promoter CpG methylation strongly negatively correlates with transcription.(67,69,71–73) Given that ESC Super Enhancers acquire CpG methylation in EpiSCs (see figure 3.1C), it follows that their associated gene promoters mirror this increase in CpG methylation. To test whether this is the case, I examined CpG methylation of ESC Super Enhancer associated gene promoters using the bisulphite-seq data provided by Drs Jouneau. The promoters of the genes associated with ESC Super Enhancers show a clear, but slight increase upon

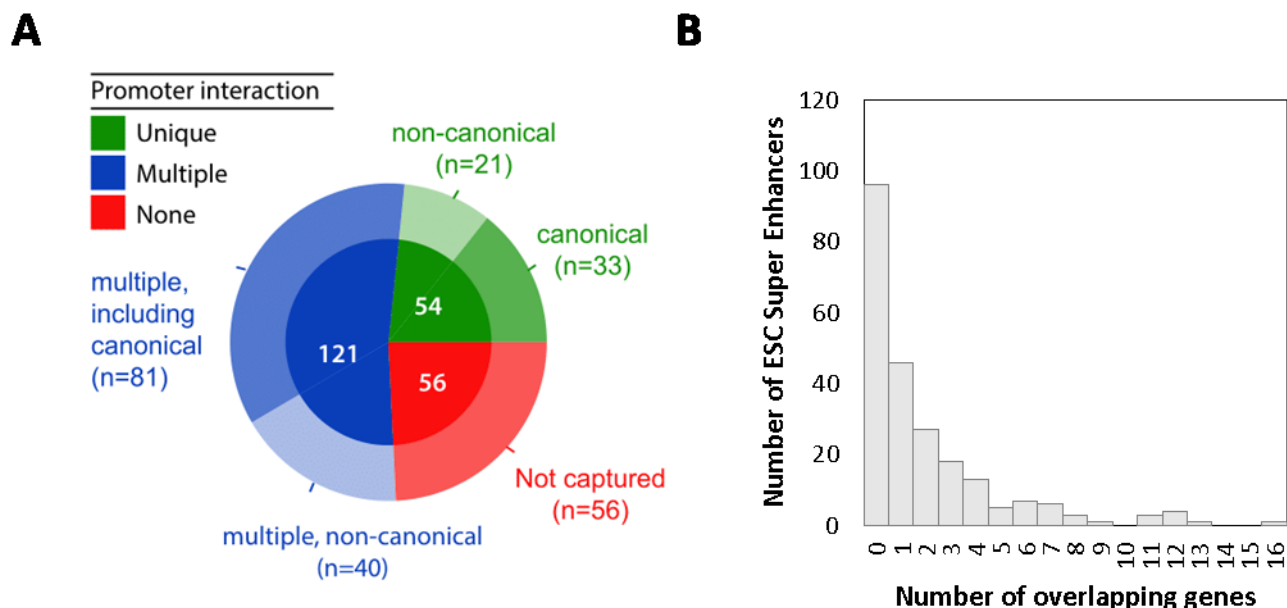


Figure 3.4 ESC Super Enhancers physically interact with their proposed associated genes.

A. Pie chart showing the number of promoters that interact with the ESC Super Enhancer assigned by Whyte et al., according to HiC-seq data analysed by Dr Wout Megchelenbrink. The inner layer shows how many of the candidate promoters associated with an enhancer region. The outer layer shows how many of those interactions were with the Super Enhancer designated by Whyte et al., (canonical) or another genomic region (non-canonical). Those with no captured interactions are coloured red. Those with a single interaction are coloured green. Those with multiple interactions are coloured blue. **B.** Histogram showing the number of gene bodies each ESC Super Enhancers overlaps with.

differentiation to EpiSCs (Mean ES-serum CpG methylation = 0.14 ± 0.02 . Mean EpiSC CpG methylation = 0.25 ± 0.03 . Paired Student t-test P-value = 1.25×10^{-9}) (figure 3.5A). However, the increase in CpG methylation appears larger at ESC Super Enhancers (Mean ES-serum CpG methylation = 0.36 ± 0.02 . Mean EpiSC CpG methylation = 0.77 ± 0.02 . Paired Student t-test P-value = 1.75×10^{-53}). A small subset of gene promoters appear hypermethylated in ESCs (serum+LIF) (figure 3.5B, left panel). Some of these genes regulate the ESC transcriptome and thus express high mRNA levels in ESCs (e.g. *Dppa3*, *Esrrb*, *Nr5a2* and *Tet1*) (table 4.1). In these cases, the ridged definition of the promoter region we used ($-900 \text{ BP} \geq \text{TSS} \geq 400 \text{ BP}$) may not capture the true promoter region. Another group of gene promoters appear similarly hypermethylated in EpiSCs (figure 3.5B, right panel). This group also contains several

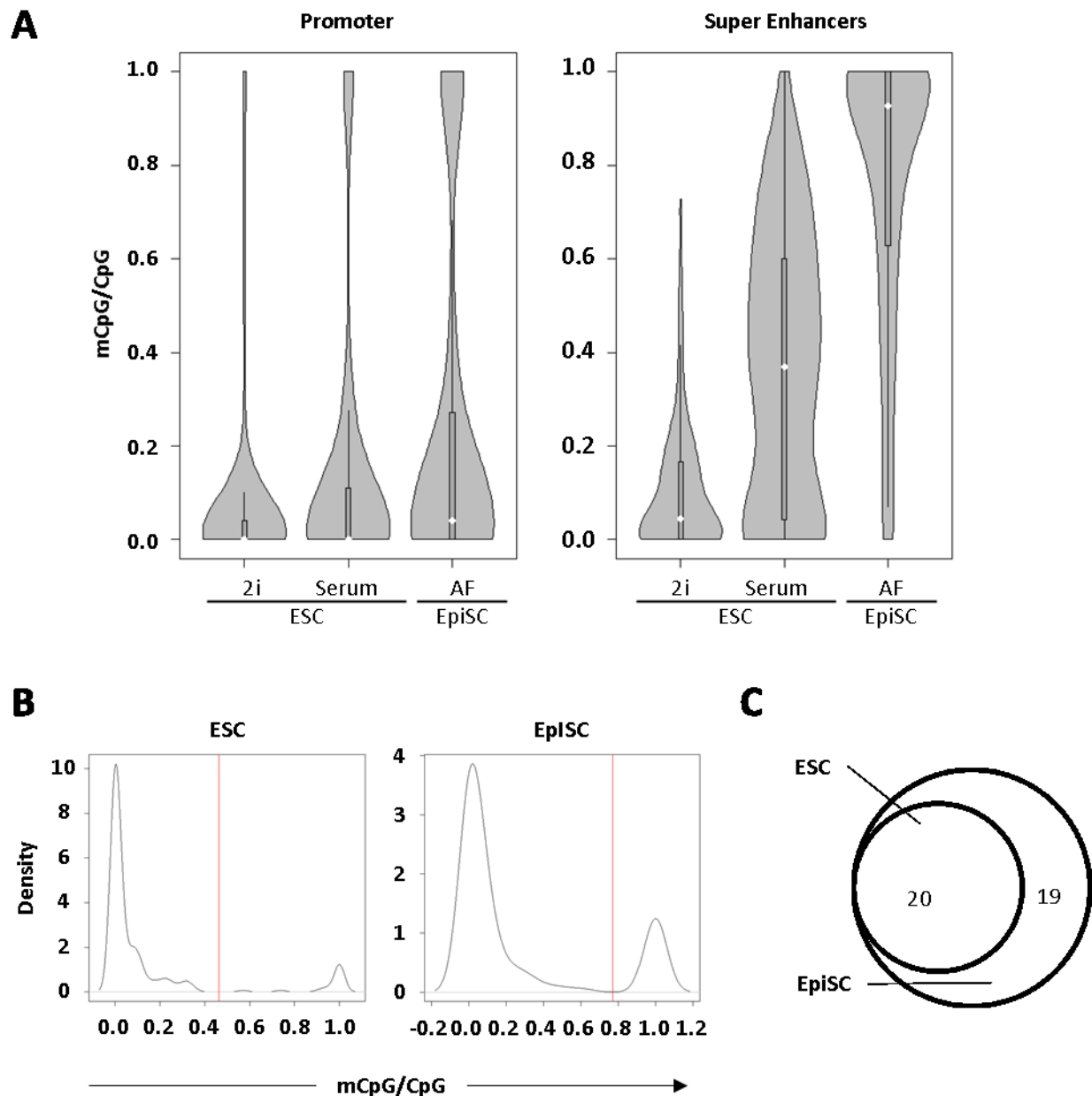


Figure 3.5 ESC Super Enhancer associated gene promoters acquire CpG methylation during pluripotency maturation.

A. Violin plots showing the mCpG/CpG at ESC Super Enhancers and their associated gene promoters during pluripotency maturation. The ESC-2i and ESC-serum data come from Habibi et al., (133) and the EpiSC data comes from Dr Alice Jouneau. **B.** Kernel density plots showing the distribution of mCpG/CpG values at the associated gene promoters of ESC Super Enhancers in ESC and EpiSCs. The mCpG/CpG values are derived from the same bisulphite-seq data plotted in **A**. The vertical red line divides the bimodal distributions into hypomethylated and hypermethylated promoters. The position of the divider was defined at the local minimum value between 0.1 and 0.9. **C.** Venn Diagram showing the overlap in hypermethylated ESC Super Enhancer associated gene promoters in ESCs and EpiSCs.

key ESC transcription factors (e.g. *Prdm14*, *Tbx3*, *Zfp42*). All of the hypermethylated ESC gene promoters also appear hypermethylated in EpiSCs (figure 3.5C) (table 4.1). In summary, EpiSCs compared to ESCs show an increase in CpG methylation at both ESC Super Enhancers and their associated gene promoters. However, the increase appears more dramatic at ESC Super Enhancers (difference in means: +0.41) than at their associated gene promoters (difference in means: +0.11). These results agree with previous findings that CpG methylation at enhancers compared to promoters more strongly negatively correlates with transcription,(78–80) implying that the most substantial epigenetic remodelling events during pluripotency maturation occur at enhancers.

Table 3.1 ESC Super Enhancer gene promoters that are CpG hypermethylated in ESCs and EpiSCs.

The genes highlighted in bold in the EpiSC column are also hypermethylated in ESCs.

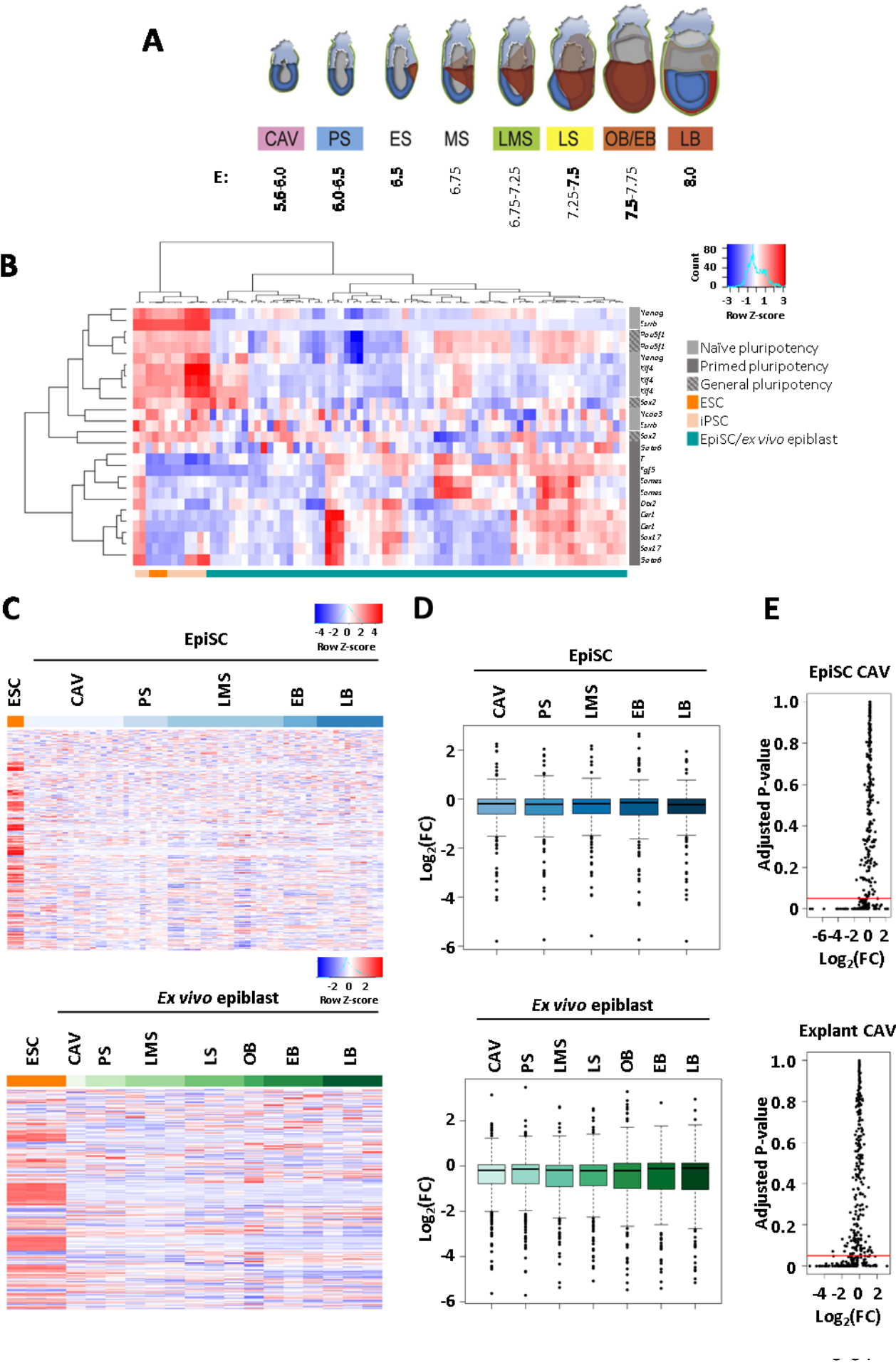
ESC	EpiSC
1700012A03Rik, 1700108F19Rik, 5830416P10Rik, <i>Dlc1</i> , <i>Dppa3</i> , <i>Esrrb</i> , <i>Ftl2</i> , <i>Glod5</i> , <i>Gm7550</i> , <i>Gm9104</i> , <i>Gpr37l1</i> , <i>Hsd17b3</i> , <i>Macf1</i> , <i>Mir145</i> , <i>Mir1929</i> , <i>Nr5a2</i> , <i>Olfr90</i> , <i>Pirt</i> , <i>Spaca7</i> , <i>Tet1</i>	1700012A03Rik, 1700108F19Rik, 5830416P10Rik, <i>Ccnd3</i> , <i>Cldn4</i> , <i>Dlc1</i> , <i>Dppa3</i> , <i>Dppa5a</i> , <i>Esrrb</i> , <i>Ftl2</i> , <i>Glod5</i> , <i>Gm16197</i> , <i>Gm1631</i> , <i>Gm7550</i> , <i>Gm9104</i> , <i>Gpa33</i> , <i>Gpr37l1</i> , <i>Gpx4</i> , <i>H2-M5</i> , <i>Hsd17b14</i> , <i>Hsd17b3</i> , <i>Lrrc2</i> , <i>Macf1</i> , <i>Mir145</i> , <i>Mir1929</i> , <i>Nphs1as</i> , <i>Nr5a2</i> , <i>Olfr90</i> , <i>Pirt</i> , <i>Prdm14</i> , <i>Pvt1</i> , <i>Rbpj</i> , <i>Spaca7</i> , <i>Tbx3</i> , <i>Tdh</i> , <i>Tet1</i> , <i>Tnip1</i> , <i>Usp7</i> , <i>Zfp42</i>

3.2.6 Post-implantation epiblast cells *in vitro* and *in vivo* downregulate transcription of ESC Super Enhancer associated genes

To determine whether EpiSCs transcribe ESC Super Enhancer associated genes, I utilised Kojima *et al*'s., extensive microarray study of EpiSCs and *ex vivo epiblast*.(5) One can derive EpiSCs across a wide range of early developmental stages (embryonic day (E) 3.5-8), (10,11,262–264) during which the embryo undergoes massive morphological change. The study aimed to determine the similarity between EpiSCs derived from these different developmental stages by comparing their transcriptomes. Kojima *et al.*, dissected epiblast

from E5.6-8 embryos (see figure 3.6A for detail), and either immediately applied them to microarray (*ex vivo* epiblast) or first cultured them to derive EpiSCs. To first test whether one can reliably distinguish ESCs and iPSCs from derived EpiSCs and *ex vivo* epiblast, I plotted the mRNA expression of factors associated with naïve (*Esrrb*, *Klf4*, *Nanog*, *Ncoa3*),^(33–36) general (*Pou5f1* (OCT4) and *Sox2*), and primed (*Cer1*, *Eomes*, *Fgf5*, *Gata6*, *Otx2*, *Sox17*, *T*)⁽³⁴⁾ pluripotency (figure 3.6B). Using Euclidian distance to cluster the probe intensity values, this gene set successfully segregates ESCs and iPSCs from EpiSCs and cognate *in vivo* epiblast tissues. The EpiSCs and epiblast broadly downregulate naïve pluripotency associated genes, and upregulate primed pluripotency associated genes. I then plotted the expression of ESC Super Enhancer associated genes in the same cell populations (figure 3.6C). Overall, EpiSCs and explanted epiblast tissues compared to ESCs and iPSCs express lower levels of ESC Super Enhancer associated gene transcripts. However, EpiSCs and epiblast do not downregulate the expression of all ESC Super Enhancer associated genes; some genes maintain or increase their mRNA expression (figures 3.6D and 3.6E). Figure 3.6D shows boxplots of the log₂ transformed fold-changes (FC) in transcription of ESC Super Enhancer associated genes between ESCs and EpiSCs/epiblast. The ESC Super Enhancer associated genes show a consistent pattern of differential expression across epiblast from each embryonic stage. Figure 3.6E shows the corresponding volcano plots of differential expression against statistical significance for the epiblast dissected from CAV. The volcano plots illustrate that in addition to those that downregulate transcription, many ESC Super Enhancer associated genes maintain their expression in epiblast, and a smaller number statistically significantly upregulate. I evaluated the enrichment of Gene Ontology (GO) terms¹⁰ in each of the upregulated, maintained, and downregulated ESC Super Enhancer associated genes (differentially expressed genes: table 3.2, GO results: table 3.3). The upregulated gene lists returned no results. This list contained only 12 (EpiSC) and 21 (*ex vivo* epiblast) genes,

¹⁰ Methods: Chapter 2.2.2.5.1



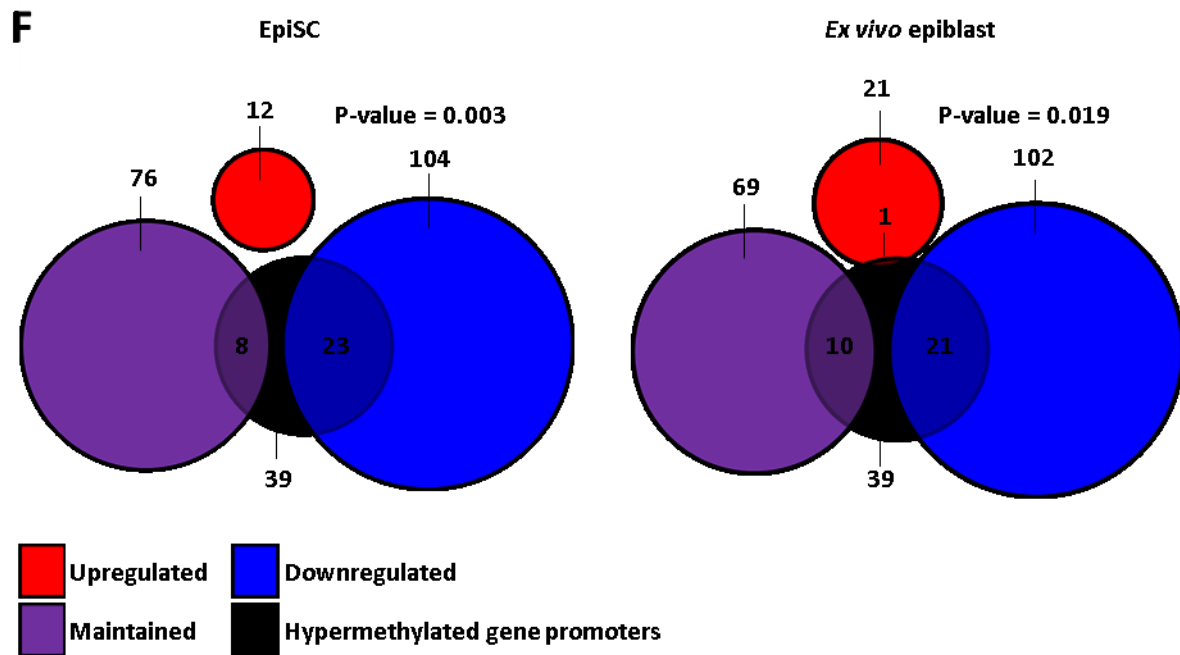


Figure 3.6 Epiblast cells, *in vitro* and *in vivo*, downregulate transcription of ESC Super Enhancer associated genes.

A. Schematic of the stages at which Kojima et al's, (5) dissected epiblast tissue for their microarray study adapted from their publication. The epiblast is coloured in blue, the mesoderm, primitive streak, and allantois in red, and the extra-embryonic mesoderm and chorion in grey. E = embryonic day. Stages: CAV = cavity, PS = pre-streak, ES = early-streak, MS = mid-streak, LMS = late mid-streak, LS = late streak, OB/EB = no bud/early bud, LB = late bud.

B. mRNA expression of transcription factors associated with naïve, primed, and general pluripotency in ESCs, iPSCs and EpiSCs and ex vivo epiblast. Normalised microarray probe intensity values are given as row-scaled Z-scores.

C. mRNA expression of ESC Super Enhancer associated genes in ESCs and EpiSCs/ex vivo epiblast. Normalised microarray probe intensity values are given as row-scaled Z-scores.

D. Boxplots of $\log_2(\text{fold-change})$ in mRNA expression between EpiSCs/ex vivo epiblast and ESCs.

E. Volcano plots of the $\log_2(\text{fold-change})$ in mRNA expression between EpiSCs/ex vivo epiblast and ESCs against their adjusted P-values. The red line indicates $P = 0.05$.

F. Venn diagrams showing the overlap between ESC Super Enhancer associated genes with hypermethylated gene promoters and differential mRNA expression in EpiSC/Epiblast E5.5 vs ESC.

Table 3.2 Super Enhancer associated genes by their differential expression in EpiSCs and ex vivo epiblast compared to ESCs.

Downregulated = $\log_2(\text{fold-change}) < -1$, Maintained = $-1 < \log_2(\text{fold-change}) < 1$, Upregulated = $1 < \log_2(\text{fold-change})$. Genes highlighted in bold are CpG hypermethylated at their promoters in EpiSCs (table 3.1, figure 3.6).

Both epiblast cell types derive from the CAV stage.

Differential expression	EpiSC	Ex vivo epiblast
Downregulated	<p><i>Alpl, Ankrd35, Bcas2, Btbd11, Cbfa2t2, Ccnd3, Cd9, Chchd10, Chd9, Cldn4, Cobl, Ctbp2, Derl3, Dlc1, Dlgap3, Dppa3, Dppa5a, Dst, Eif4a2, Enah, Esrrb, Etl4, Fgf4, Fnbp1, Gadd45a, Gli2, Glod5, Gm1631, Gpa33, Gpx4, H2-M5, Hsd17b14, Hspb8, Itpk1, Jam2, Kat6b, Kazn, Klf2, Klf4, Klf5, Klf9, Lamc2, Lefty2, Lrrc2, Manba, Mapt, Med13l, Mkrn1, Mreg, Msi2, Mybl2, Nanog, Neat1, Nr5a2, Phc1, Pim1, Pipox, Pirt, Pitpnc1, Ppp2r5c, Prdm14, Prrc2b, Pum1, Pvt1, Ranbp17, Rbfox2, Rbpms, Rhof, Sall1, Sall4, Sema4b, Sgk1, Sik1, Smarcd1, Sox2, Spry4, Ssr2, Stoml1, Sulf2, Tbx3, Tcf15, Tdh, Tet1, Tet2, Tfcp2l1, Tgif1, Tmem131, Tnip1, Tns3, Trim71, Ube2s, Ubtf, Uck2, Upp1, Uri1, Usp48, Usp7, Utf1, Zbtb45, Zfhx2, Zfml, Zfp281, Zfp42, Zfp710</i></p>	<p><i>Alpl, Ankrd35, Btbd11, Capns1, Cbfa2t2, Ccnd3, Cd9, Chchd10, Chd9, Cobl, Ctbp2, Derl3, Dlc1, Dlgap3, Dppa3, Dppa5a, Dst, Eif4a2, Enah, Epha2, Esrrb, Fbxo36, Fgf4, Fnbp1, Gadd45a, Gbx2, Gli2, Glod5, Gm1631, Gpa33, Gpx4, H2-M5, Hs6st1, Hsd17b14, Hspb8, Ift52, Inhbb, Jam2, Kat6b, Kazn, Klf13, Klf2, Klf4, Klf5, Klf9, Lamc2, Lefty1, Lefty2, Lrrc2, Manba, Mapt, Med13l, Mreg, Mybl2, Nanog, Nfkb1a, Nr5a2, Phc1, Pim1, Pirt, Pitpnc1, Pou5f1, Ppp2r5c, Prdm14, Prrc2b, Pvt1, Rara, Rbpms, Rhof, Sall1, Sall4, Sema4b, Sik1, Slc2a3, Slc6a6, Smarcd1, Socs3, Sox2, Spry4, Ssr2, Stoml1, Sulf2, Tbx3, Tdh, Tead1, Tet2, Tfcp2l1, Tgif1, Tnip1, Tns3, Trim71, Tsc22d1, Uck2, Upp1, Uri1, Usp48, Usp7, Zbtb34, Zbtb45, Zfp42, Zfp704, Zfp710</i></p>
Maintained	<p>1700012A03Rik, 5830416P10Rik, Abhd2, Agtrap, B330016D10Rik, Brd1, C2cd5, Capns1, Cenpv, Col18a1, Ctif, Dmrt1, Dusp1, Elovl6, Enc1, Epha2, F2rl1, Fam134b, Fam53a, Fbxo36, Gbx2, Gm7550, Gpr37l1, Gpt2, Grik3, Gtf3c6, Hs6st1, Hsd17b3, Id1, Ier2, Ift52, Inhbb, Irf2bpl, Kank4, Kctd16, Kif13b, Klf13, Klf3, Ldhb, Lefty1, Macf1, Mbip, Mcl1, Mesdc2, Mllt6, Mycn, Nav2, Ndfip1, Nfkb1a, Olfr90, Opa1, Otx2, Pou5f1, Ppp1r1a, Pura, Qk, Rara, Rbpj, Rpap3, Rpl14, Ski, Slc2a1, Slc2a3, Slc6a6, Smad7, Socs3, Soga2, Spaca7, Spry2, Tead1, Tmem220, Tnfsf12, Tsc22d1, Vdac1, Zbtb34, Zfp704</p>	<p>1700012A03Rik, 5830416P10Rik, Abhd2, Agtrap, Ahi1, Amigo2, B330016D10Rik, Bcas2, C2cd5, Camk2n1, Cenpv, Cldn4, Col18a1, Ctif, Dmrt1, Dusp1, Elovl6, Enc1, F2rl1, Fam134b, Fam53a, Gm7550, Gpr37l1, Gpt2, Grik3, Hsd17b3, Idh2, Ier2, Irf2bpl, Itpk1, Kank4, Kctd16, Klf3, Ldhb, Mcl1, Mesdc2, Mllt6, Mycn, Ndfip1, Neat1, Olfr90, Opa1, Pipox, Ppp1r1a, Pura, Qk, Ranbp17, Rbfox2, Rbpj, Reep3, Rpap3, Sgk1, Slc2a1, Smad7, Soga2, Spaca7, Spry2, Tcf15, Tet1, Tmem131, Tmem220, Tnfsf12, Trak1, Ube2s, Utf1, Vdac1, Zfhx2, Zfml, Zfp281</p>
Upregulated	<p>6430573F11Rik, Ahi1, Amigo2, Camk2n1, Ddit4, Hmg20a, Idh2, Igfbp2, Kras, Reep3, Tmem2, Trak1</p>	<p>6430573F11Rik, Brd1, Ddit4, Etl4, Gtf3c6, Hmg20a, Id1, Igfbp2, Kif13b, Kras, Macf1, Mbip, Mkrn1, Msi2, Nav2, Otx2, Pum1, Rpl14, Ski, Tmem2, Ubtf</p>

Table 3.3 Gene Ontology terms enriched for of ESC Super Enhancer associated genes.

Gene Ontology (G) terms enriched ESC Super Enhancer associated genes that are in downregulated (blue) and upregulated (pink) in EpiSCs/ex vivo epiblast (CAV, E5.5) compared to ESCs. The values represent the $-10 \times \log_{10}(P\text{-value})$. GO was performed using PANTHER DB(243) PANTHER Overrepresentation Test. P-values were corrected for multiple hypothesis testing using the Bonferroni method.

Gene Ontology Term	EpiSC	Ex vivo epiblast
cellular response to leukemia inhibitory factor (GO:1990830)	13.85	13.78
response to leukemia inhibitory factor (GO:1990823)	13.85	13.78
single-organism developmental process (GO:0044767)	6.62	6.99
neural tube development (GO:0021915)	2.68	6.07
cellular response to cytokine stimulus (GO:0071345)	4.28	5.10
stem cell population maintenance (GO:0019827)	5.05	5.00
cellular developmental process (GO:0048869)	3.85	4.25
response to cytokine (GO:0034097)	3.29	4.04
morphogenesis of embryonic epithelium (GO:0016331)	1.84	4.02
neural tube formation (GO:0001841)	1.66	4.00
cell differentiation (GO:0030154)	3.57	3.96
embryonic epithelial tube formation (GO:0001838)		3.43
neural tube closure (GO:0001843)	2.17	3.38
tube closure (GO:0060606)	2.14	3.35
epithelial tube formation (GO:0072175)		3.29
primary neural tube formation (GO:0014020)	1.99	3.17
tube formation (GO:0035148)		2.97
animal organ development (GO:0048513)		2.50
endodermal cell fate specification (GO:0001714)		1.51
notochord development (GO:0030903)		1.44
regulation of activin receptor signaling pathway (GO:0032925)	1.88	

therefore would have limited power to detect significantly enriched terms. The downregulated gene lists from both EpiSC and ex vivo epiblast returned numerous developmental pathways, but interestingly appear most statistically significantly enriched for two GO terms related to LIF signalling. The maintained gene list returned only one statistically significantly enriched GO term, pertaining to Activin receptor signalling. ESCs and EpiSCs respectively rely on LIF and Activin to maintain self-renewal in culture. Finally, to link the differential expression of ESC Super Enhancer associated genes in epiblast to their epigenetic regulation, I compared the overlap of upregulated, maintained, and downregulated genes with the hypermethylated gene

promoters described in chapter 3.2.5 (figure 3.6E, table 3.2). In both EpiSC and *ex vivo* epiblast derived from CAV stage embryos, only the downregulated genes show a statistically significant overlap with hypermethylated gene promoters (EpiSC: P-value=0.003. *Ex vivo* epiblast: P-value=0.019). In summary, many ESC Super Enhancer associated genes reduce expression in epiblast both *in vitro* and *in vivo*. Increased gene promoter CpG methylation accompanies this change. Some of these genes contribute to LIF signalling, which ESCs require in culture. However, other ESC Super Enhancer associated genes retain or even increase expression. Some of these genes contribute to Activin signalling, which EpiSCs rely on in culture. Together these observations suggest that ESC Super Enhancers regulate transcription of a heterogeneous group of self-renewal and pluripotency factors, not just those associated with naïve pluripotency.

3.2.7 Genes associated with ESC Super Enhancer compared to ESC Typical

Enhancers appear more sensitive to the loss of naïve pluripotency factors

Whyte *et al* previously found that upon shRNA knock down of MED1 or OCT4 over 3-5 days, ESCs broadly downregulate ESC Super Enhancer associated genes, but not Typical Enhancer associated genes or house-keeping genes. I therefore posited that pluripotency maturation (i.e. the loss of naïve pluripotency genes and expression of general and primed post-implantation pluripotency factors illustrated in figure 3.7B) selectively deregulates the transcription of ESC Super Enhancer associated genes but not Typical Enhancer associated genes or housekeeping genes. I used the PaGenBase database(265) to source the house-keeping gene list. I returned to Kojima *et al*'s microarray study to compare their expression, along with all remaining probes on the array, in naïve and primed pluripotent cells (figure 3.7A). EpiSCs and *ex vivo* epiblast tissues consistently downregulate ESC Super Enhancer associated genes, but not the other gene sets. Conversely, the same four gene sets show a reciprocal expression pattern in ESCs (figure 3.7B) and iPSCs (figure 3.7C) relative to more differentiated MEFs. ESCs/iPSCs exhibit higher transcription of ESC Super Enhancer associated genes, but not the other gene sets. Together, these results show that ESCs

transcribe higher levels of ESC Super Enhancer associated genes compared to primed pluripotent and more differentiated cell types.

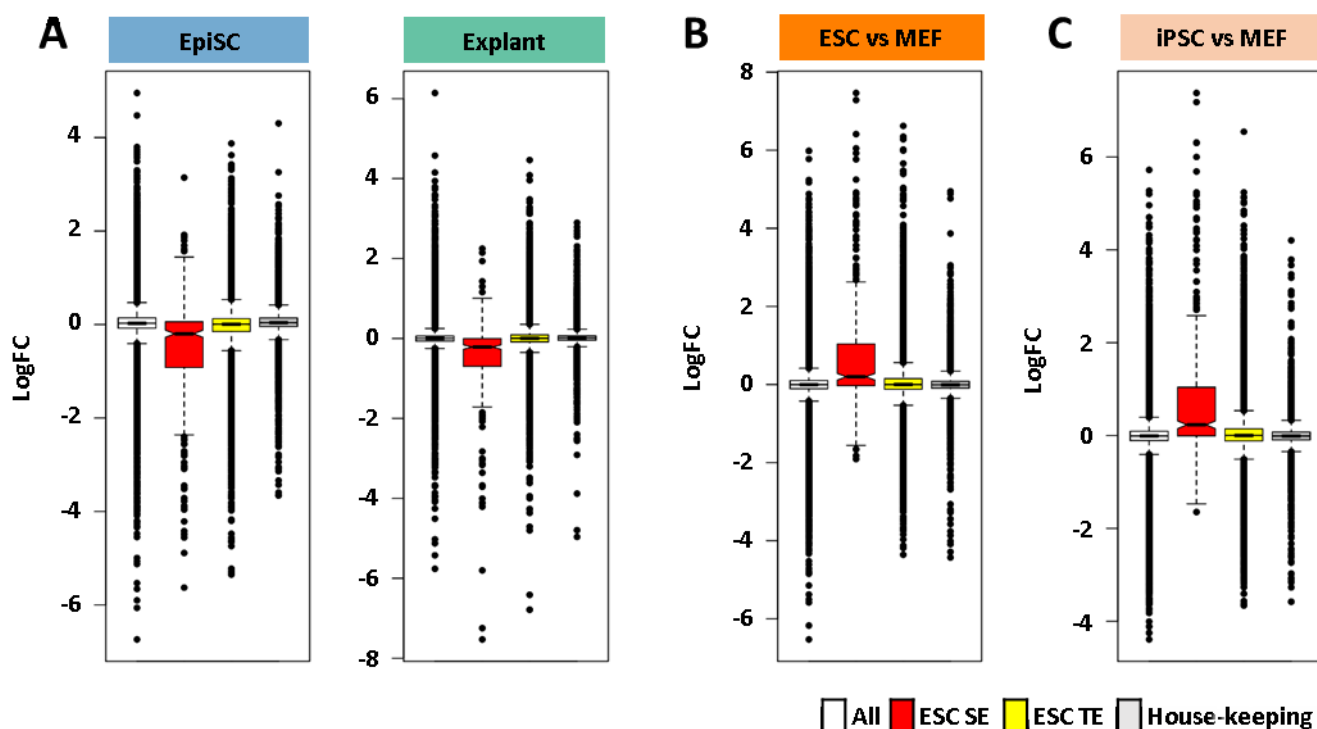


Figure 3.7 ESC Super Enhancer associated genes selectively deregulate during pluripotency maturation.

ESC Super Enhancer associated genes, but not Typical Enhancer associated or house-keeping genes selectively deregulate as pluripotency matures. This figure shows boxplots of $\log_2(\text{fold-change})$ in mRNA expression of these three gene sets, plus all remaining probes on the microarray, in: **A.** EpiSCs and ex vivo epiblast derived from the CAV stage compared to ESCs; **B.** MEFs compared to ESCs; **C.** MEFs compared to iPSCs. House-keeping genes were defined using the PaGenBase.

3.3 Conclusions

In this chapter I aimed to determine whether the ESC Super Enhancers defined by Whyte *et al.*, regulate both the naïve pluripotency of ESCs and the primed pluripotency of EpiSCs. I profiled three indicators of enhancer activity to do so: CpG methylation, transcription factor binding, and transcription of associated genes.

Overall, ESC Super Enhancers maintain a heterogeneous but relatively hypomethylated state in ESCs and become hypermethylated in EpiSCs (figure 3.1C). In contrast, Super Enhancers

defined in differentiated cells appear hypermethylated in ESCs (figure 3.2C). Given that CpG methylation of enhancers strongly negatively correlates with their activity,(74–80) this suggests that ESC Super Enhancers lose activity in EpiSCs. ESC Super Enhancer associated gene promoters acquire CpG methylation in EpiSCs, which further evidences their transcriptional repression.(67–73) However, Super Enhancers display a much more dramatic shift in CpG methylation (figure 3.5A), in agreement with the published literature.(78–80)

A closer examination of CpG methylation at ESC Super Enhancers reveals discrete blocks of unmethylated and methylated CpGs (figure 3.3A). When defining Super Enhancers Whyte *et al.*, joined together any enhancers within 12 KB of each other. Most ESC Super Enhancers therefore consist of multiple constituent enhancers and interstitial subregions. In ESCs, transcription factors (e.g. OCT4, SOX2, NANOG, ESRRB, KLF4) bind to enhancers in multi-protein complexes, (17,18) recruit chromatin modifiers (e.g. the Mediator and Cohesin complexes, and the histone deacetylase P300), (180,249) and provide an anchor for components of the transcriptional machinery.(249) Therefore, enhancers form dense protein binding hubs. Transcription factor binding clusters at the unmethylated Super Enhancer subregions, whereas the interstitial subregions appear relatively barren (figure 3.3D). This suggests the unmethylated regions correspond to the constituent enhancer units within ESC Super Enhancers.

ESC Super Enhancer associated genes tend to decrease transcription as pluripotency matures both *in vitro* and *in vivo* (figure 3.7C, 3.7D, and 3.7E). In contrast, ESC Typical Enhancer associated genes show no clear pattern of expression (figure 3.8A). However, some ESC Super Enhancer associated genes continue or even increase their expression (figure 3.7D and 3.7E). The genes downregulated in EpiSCs and *ex vivo* epiblast associate with the LIF pathway, a key driver of naïve pluripotency(8,9) (table 3.3). In contrast, the maintained genes associate with Activin signalling, a key driver of primed pluripotency(10,11). Furthermore, some ESC Super Enhancer associated gene promoters appear almost completely methylated in EpiSCs (figure 3.6B). These genes include key naïve pluripotency

factors such as *Esrrb*, *Prdm14*, and *Tbx3* (table 3.1). These results suggest that the genes associated with ESC Super Enhancers represent both naïve and primed pluripotency.

Collectively, these results evidence that overall, ESC Super Enhancers pertain to the naïve pluripotency of ESCs. However, that some ESC Super Enhancer associated genes maintain or increase their expression in epiblast suggests a more complex model than a complete decommissioning of these regulatory regions.

4 RESULTS II

SELECTIVE DECOMMISSIONING OF ESC-SPECIFIC 'SUPER ENHANCELETS' IN EPISCS AND IN POST- IMPLANTATION EPIBLAST IN VIVO

4.1 Introduction

The experiments described in Chapter 3: Results I demonstrate that total CpG methylation at ESC Super Enhancers increases as pluripotency matures. In tandem, their associated genes slightly but consistently increase CpG methylation at their promoters, and broadly reduce transcription. Initially, these observations led us to conclude that ESC Super Enhancers decommission in EpiSCs. However, a more complex picture emerged upon closer examination of the transcription of ESC Super Enhancer associated genes in EpiSCs. While many of these genes reduce transcription in EpiSCs, some maintain or even increase transcription. The genes that reduce transcription contain members of the LIF-signalling pathway, whereas the genes that maintain/increase transcription contain members of the Activin signalling pathway. These pathways respectively drive naïve ESC and primed EpiSC pluripotency. This suggests that the Super Enhancers Whyte *et al.*,(179) defined in ESCs might contribute to the regulation of both naïve and primed pluripotency genes.

ESC Super Enhancers contain one or more constituent enhancers, henceforth termed Super Enhancelets, to which pluripotency regulatory proteins bind at high density. ESC-serum bisulphite-seq data analysed using a Hidden Markov Model (HMM) demarcates blocks of consecutively unmethylated CpGs that correspond to the Super Enhancelets. The remaining portion of the Super Enhancer, henceforth termed the interstitial subregions, appear as blocks of consecutively methylated CpGs. In this chapter, I build on our previous findings (i.e. that ESC Super Enhancer associated genes display heterogeneity in their differential expression in epiblast derived cells) by considering the epigenetic and transcriptional profiles of ESC

Super Enhancelets during pluripotency maturation *in vitro* and *in vivo*. I aim to determine whether the patterns of differential gene expression of ESC Super Enhancer associated genes in ESCs and EpiSCs relate to differing properties of their Super Enhancers.

4.2 Results

4.2.1 ESC Super Enhancers acquire CpG methylation in distinct patterns in EpiSCs

In the previous chapter I found that ESC Super Enhancers increase their total CpG methylation during the transition from naïve (ESCs) to primed (EpiSCs) pluripotency. As CpG methylation at enhancers strongly negatively correlates with activity(74–80), this implies that these enhancers decommission in EpiSCs. However, while many of their associated genes reduce transcription in EpiSCs, some maintain or increase in expression (figure 3.6C-E). To further investigate the fate of ESC Super Enhancers in EpiSCs, we looked at the CpG methylation profiles of individual ESC Super Enhancelets in ESCs and EpiSCs (figure 4.1A). Drs Alice and Luc Jouneau provided bisulphite-seq data to do this. Figure 4.1A shows the methylation status of each CpG within the *Klf4*-, *Klf13*-, and *Lefty1*-associated ESC Super Enhancers. Drs Jouneau used a Hidden Markov Model (HMM) to designate each CpG as ‘methylated’ or ‘unmethylated’.¹¹ The y-axis displays the β methylation scores, which represent the statistical evaluation of methylation between 0 (unmethylated) and 1 (methylated). CpGs designated as unmethylated have had their β methylation score subtracted from 1 and multiplied by -1 to place their value between -1 and 0. The coloured segments indicate the position of Super Enhancelets, which we previously classified in ESCs grown in serum+LIF as overall unmethylated subregions. While some Super Enhancelets do acquire CpG methylation in EpiSCs, others maintain an unmethylated state. We termed these Super Enhancelets Differentially Methylated (DM; pink) and Persistently Unmethylated (PU; green) respectively. The interstitial regions (grey) between the Super Enhancelets that appear methylated in both

¹¹ Methods: Chapter 2.2.2.4.1.2

cell types we termed Persistently Methylated (PM). ESC Super Enhancers display three patterns of CpG methylation at Super Enhancelets in ESCs and EpiSCs:

1. All DM (e.g. *Klf4*-associated ESC Super Enhancer);
2. Some PU and some DM (e.g. *Klf13*-associated ESC Super Enhancer);
3. All PU (e.g. *Lefty1*-associated ESC Super Enhancer).

If we divide each ESC Super Enhancer into discrete units of Super Enhancelets and PM subregions, we see that Super Enhancelets form 43% of all units (figure 4.1B). If we divide the Super Enhancelets into two groups based on their methylation status in EpiSCs we find the PU and DM respectively comprise 26% (11% of total units) and 74% (32% of total units), i.e. most Super Enhancelets become hypermethylated in EpiSCs.

The HMM Drs Jouneau used to designate a given CpG as methylated or unmethylated infers that probability from the methylation status of neighbouring CpGs. Therefore it is biased to predict blocks of consecutively unmethylated or methylated CpGs. To assess the variation in CpG methylation within PU, DM and PM subregions as pluripotency matures, I assessed the mean of the ratio of mCpG to CpG sequencing reads in ESCs and EpiSCs (figure 4.1C). Drs Jouneau filtered out CpGs with less than 7 reads. In ESCs cultured in both 2i+LIF and serum+LIF, the vast majority of PU and DM show a mean mCpG/CpG of 0. In both culture conditions the DM display more outliers than the PU mean mCpG/CpG (i.e. mean mCpG/CpG $> 1.5 \times$ interquartile range). This could indicate greater heterogeneity in CpG methylation at DM and/or less efficient active demethylation. In the EpiSCs the PU and DM display starkly different fates, with the PU totally unmethylated and the DM almost totally methylated. The PM display a stepwise increase in mean mCpG/CpG as pluripotency progresses. The PM may reflect the increase in CpG methylation observed in pre-implantation blastocysts,(123,124) whereas a targeted epigenetic remodelling event may cause the hypermethylation of the DM.

As figure 4.1A shows, PU, DM, and PM subregions vary in length, and thus may contain different numbers and/or normalised fraction of CpGs. mCpG readily deaminates to thymine

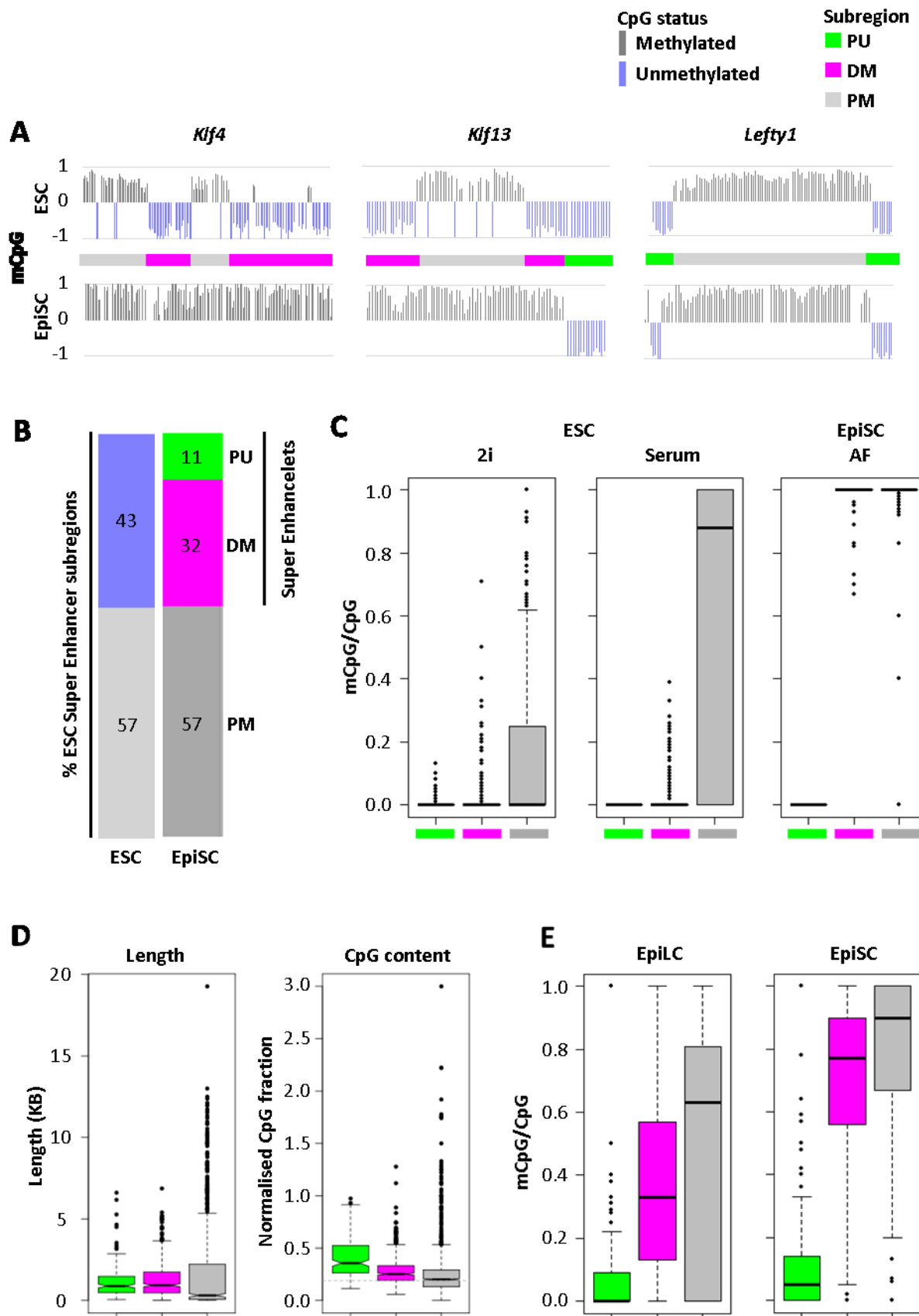
causing underrepresentation of CpG dinucleotides throughout the genome. Therefore, regions containing higher CpG content are under selective evolutionary pressure either to retain CpGs or preserve the DNA sequence of the whole region. The latter would indicate a functional role for that genomic region, while the former specifically points to regulatory CpG methylation. To assess potential differences, I assessed the length and CpG content of these subregions (figure 4.1D). The Super Enhancelets show similar distributions in terms of length, but the PM appear much more variable. The original definition of ESC Super Enhancers joined any high density MED1 binding sites within 12 KB of one another. Therefore the length of the PM depends on the density of MED1 binding sites within a given genomic region. Interestingly, the normalised CpG fraction¹² appears statistically significantly higher in PU compared to DM. This implies that PU compared to DM have been subjected to stronger selective pressure to conserve either their CpG content or their entire DNA sequence. Several possible explanations immediately present themselves: PU could act as regulatory genomic regions in multiple cell types and therefore occur in a CpG unmethylated state more frequently than DM, preventing mCpG deamination; PU compared to DM could drive the expression of genes more critical to ESCs, thus leading to stronger selection against mutation in these regions; or epigenetic mechanisms of regulation could play a larger role at PU compared to DM.

To examine whether CpG methylation acquisition during pluripotency maturation occurs in a gradual, step-wise manner, I attempted to replicate the CpG methylation profiles of ESC Super Enhancer subregions in an independent Epiblast-Like Cell (EpiLC) and EpiSC bisulphite-seq,(38) processed by Drs Jouneau (figure 4.1E). The EpiLC represent a transitional stage between the ESC and EpiSC, as ESCs incubated for 48 hours in EpiSC media (N2B27+Activin+Fgf).(38) The PU appear hypomethylated compared to the DM and PM in both cell types. The DM appear intermediately methylated compared to the PU and PM. Both the DM and PM show CpG hypermethylation in the EpiSCs compared to the EpiLCs. This confirms the increase in CpG methylation of DM, and not PU seen in figure 4.1C, and that

¹² Methods: Chapter 2.2.2.1

CpG methylation increases in a step-wise manner with pluripotency maturation. To directly compare the mCpG/CpG scores across ESC Super Enhancer subregions in ESCs,(74,133) EpiLCs(38) and EpiSCs,(38) I next assessed these data a heatmap (figure 4.1F). I clustered the cell types by Pearson correlation and the subregions by Euclidian distances. These data include two independently generated ESC (serum+LIF) and EpiSC samples clustered. The duplicate samples cluster together as expected. The ESC (2i+LIF) and EpiLC datasets cluster with the ESC (serum+LIF) datasets. This confirms that the EpiLC cells represent an intermediate between the ESC (2i+LIF) and the EpiSCs, and that despite the markedly different DNA methylation profiles ingenerated by 2i+LIF and serum+LIF culture conditions, the ESC samples behave similarly to one another. The ESC Super Enhancer subregions form almost perfectly distinct PU, DM, and PM clusters. One ought to expect this to some extent as the Super Enhancelet definitions come from two of the contributing data sets. However, that the second ESC (serum+LIF) and EpiSC samples so closely resemble the ESC (serum+LIF) and EpiSC datasets used to define these regions, implies consistent epigenetic regulation of these subregions in these two states of pluripotency. Finally, I validated the methylation profiles of ESC Super Enhancer subregions in ESCs and EpiSCs using an independent DIP-seq dataset(129) targeting 5-mC (figure 4.1G).¹³ As observed with bisulphite-seq, in ESCs both DM and PU Super Enhancelets appear hypomethylated compared to PM subregions, but only the DM Super Enhancelets acquire CpG methylation in the EpiSCs. As CpG methylation strongly negatively correlates with enhancer activity, this implies that DM, but not PU Super Enhancelets decommission as pluripotency matures.

¹³ Methods: Chapter 2.2.2.4.1.4



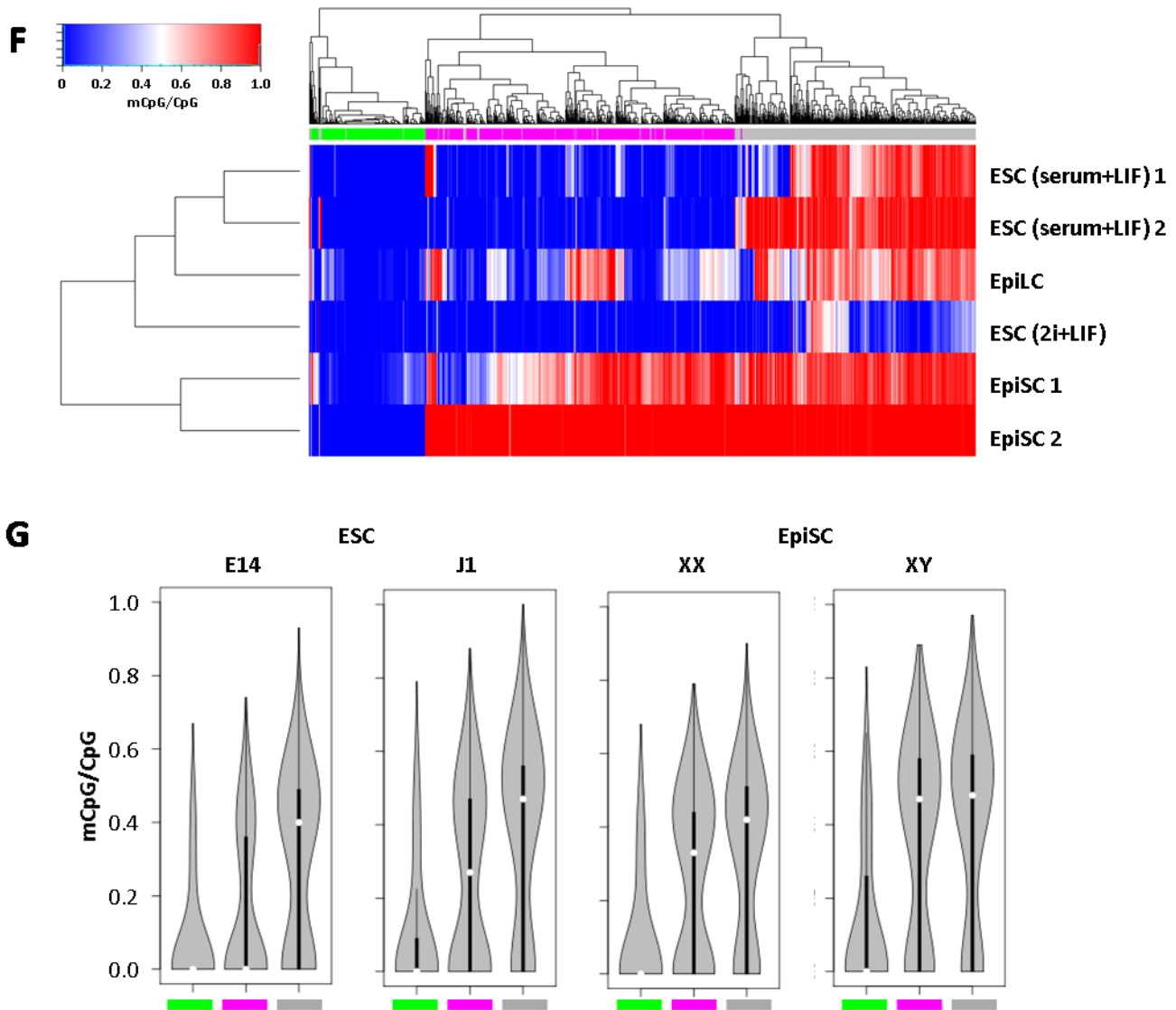


Figure 4.1 ESC Super Enhancers acquire CpG methylation in distinct patterns in EpiSCs.

A. The methylation status of each CpG within the *Klf4*-, *Klf13*-, and *Lefty1*-associated ESC Super Enhancers in ESCs (serum+LIF) (74,133) and EpiSCs (unpublished data courtesy of Dr Alice Jouneau), as assayed by bisulphite-seq (data provided by Drs Alice and Luc Jouneau). Drs Jouneau used a Hidden Markov Model (HMM) to designate each CpG as 'methylated' or 'unmethylated'. Methylated CpGs are coloured grey and unmethylated CpGs are coloured blue. The y-axis displays the β methylation scores. CpGs designated as unmethylated have had their β methylation score subtracted from 1 and multiplied by -1 to place their value between -1 and 0. The magenta (Differentially Methylated or DM) and green (Persistently Unmethylated or PU) coloured segments indicate the locations of Super Enhancers. **B.** The proportion of different ESC Super Enhancer subregions. In ESCs 43% of subregions are unmethylated constituent enhancer units and 57% are interstitial methylated regions. In EpiSCs, the unmethylated constituent enhancer units divide into PU (11%) and DM (32%). **C.** Boxplots showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in ESCs grown in 2i+LIF and serum+LIF, (74,133) plus EpiSCs grown in N2B27 with Activin and bFGF (AF) (unpublished data courtesy of Dr Alice Jouneau).

D. Boxplots showing the length (left panel) and normalised CpG fraction (right panel) of PU (green), DM (magenta) and PM (grey) Super Enhancer subregions. **E.** Boxplots showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in EpiLCs and EpiSCs. (38) **F.** Heatmap showing the proportion of methylated CpGs in ESCs grown in 2i+LIF and serum+LIF, (74,133) EpiLCs, and EpiSCs. (38) The cell types were clusters by Pearson correlation and the subregions by Euclidian distance. **G.** Violin plots of MeDIP-seq data(129) showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions, in ESC and EpiSC cell lines.

4.2.2 ESC Super Enhancers selectively acquire CpG methylation in epiblast *in vivo*

During pre-implantation development embryos undergo global DNA demethylation reaching the lowest levels by the early blastocyst stage (E3.5). After the blastocyst implants into the uterine wall, DNMT3A and DNMT3B re-establish CpG methylation in the ICM.(136,266) The global CpG hypomethylation and hypermethylation of pre-implantation and post-implantation embryos respectively resemble the epigenetic states of ESCs and EpiSCs.(131–133,267) Therefore, *in vivo* pre- and post-implantation embryos should recapitulate the changes in CpG methylation observed at ESC-specific Super Enhancer subregions *in vitro*.

To test this hypothesis I utilised independent whole genome (WGBS) (unpublished data provided by Dr Peter Rugg-Gunn) and reduced representation bisulphite-seq (RRBS)(266) data (processed by Drs Alice and Luc Jouneau). Figure 4.2A shows the mCpG/CpG in E3.5 and E4.5 ICM and E5.5 and E6.5 Epiblast from the WGBS. As previously seen in ESCs and EpiSCs, at each time-point *in vivo* the PU appear almost totally unmethylated. The DM show similar hypomethylation to the PU in E3.5 ICM, but progressively accumulate CpG methylation from E4.5. The PM show low levels of CpG methylation at E3.5 and E4.5, similar to those observed in ESCs cultured in 2i+LIF (figure 4.1C). From E5.5 these regions progressively acquire high levels of CpG methylation. The RRBS data (figure 4.2B) appears similar to the WGBS (figure 4.2A), although all three subregions appear slightly hypermethylated in the RRBS compared to the WGBS at E3.5. This hypermethylation could result from differences in genetic background of the mice. There are also feasible technical reasons for the discrepancy.

RRBS employs restriction enzyme that cuts at 5-mC but not 5-C sites to digest the genome followed by size selection. Therefore, the sequencing library will inherently bias towards 5-mC fragments, which could result in an artificial appearance of CpG hypermethylation. Dr Rugg-Gunn uses Bismark(268) instead of the HMM method employed by Dr Jouneau. Therefore, differences in the underlying statistical and mathematical assumptions of these methods may contribute to the observed differences in CpG methylation. Regardless, the DM and PM subregions acquire CpG methylation at E6.5 in a similar pattern. Interestingly, E6.5 and E7.5 epiblast shows the same levels of CpG methylation. This may indicate that CpG methylation acquisition slows down and/or plateaus as the epiblast matures. In summary, the *in vivo* ICM and epiblast does recapitulate the selective acquisition of CpG methylation at DM and not PU ESC Super Enhancelets.

4.2.3 ESC Super Enhancelets display distinct methylation patterns in non-pluripotent cell types

Thus far we have observed that the PU appear largely CpG unmethylated in pluripotent cells both *in vitro* and *in vivo*. Conversely, the DM acquire CpG methylation as pluripotency matures. As CpG methylation of enhancers strongly negatively correlates with activity(74–80), this potentially reflects that the DM promote the naïve pluripotency network, whereas the PU promotes both the naïve and primed pluripotency network. However, one can't assume specificity to pluripotency without comparison to non-pluripotent cell types. At E4.5 the murine blastocyst contains three lineages: ICM, extra-embryonic trophectoderm (TE), and primitive endoderm (PrE). In addition to the pluripotent ESCs sourced from the ICM, one can derive trophoblast stem cells (TS)(269) from the TE and extra-embryonic ectoderm (XEN)(270) stem cells from the PrE. To examine whether the patterns of CpG methylation at PU and DM during pluripotency maturation are specific to pluripotency, I analysed CpG methylation of these Super Enhancelets in extra-embryonic and more differentiated cell types, using a combination of publicly available and privately shared whole genome methylation data.

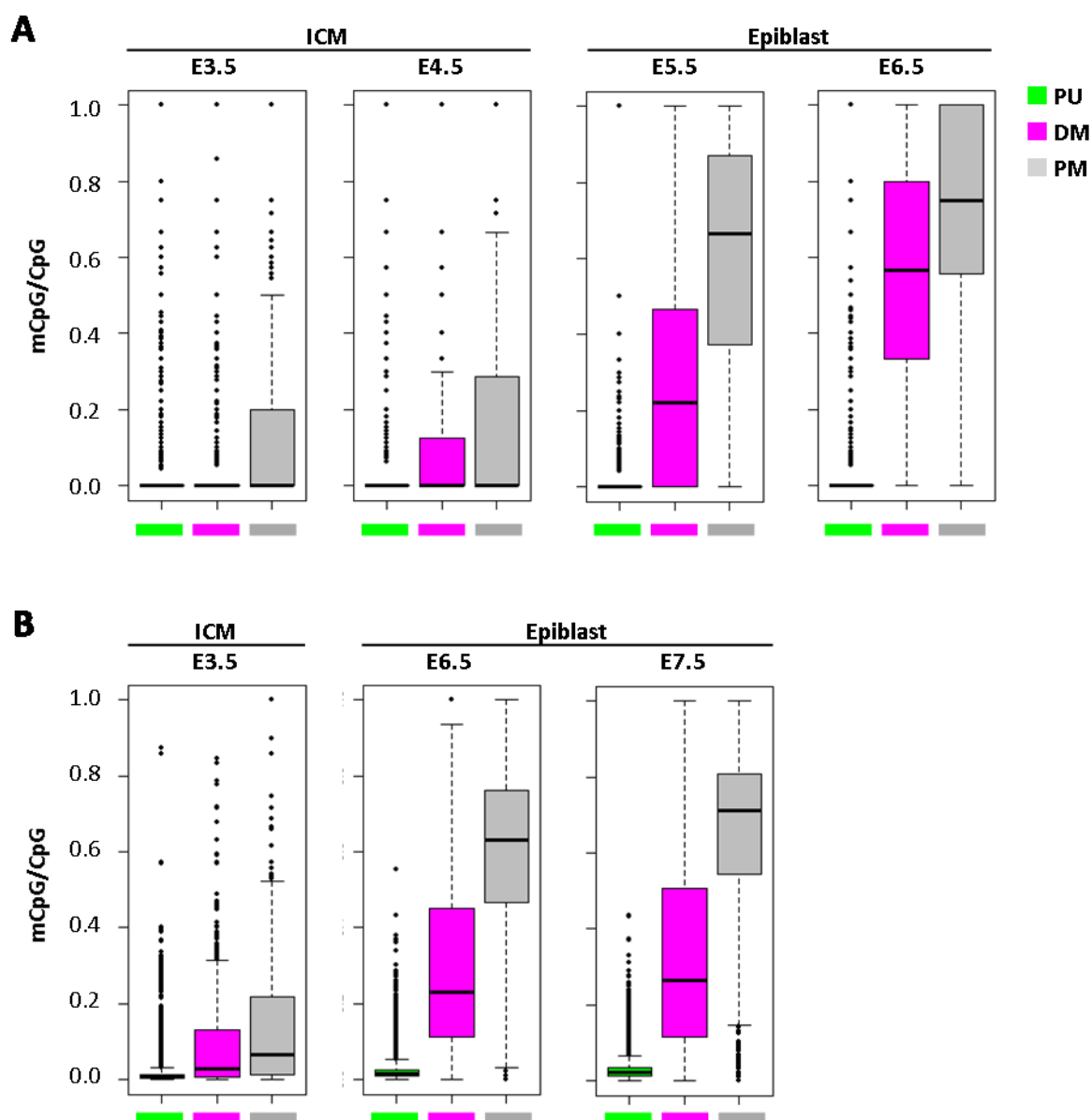


Figure 4.2 ESC Super Enhancers acquire CpG methylation in distinct patterns in epiblast in vivo.

Boxplots of bisulphite-seq data showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in tissue derived from ICM and Epiblast. E = embryonic day. **A.** Whole genome bisulphite-seq from Dr Peter Rugg-Gunn's group (unpublished data). **B.** Reduced Representation bisulphite-seq (RRBS) from Prof Alexander Meissner's group. (346)

Firstly, I used publicly available MeDIP-seq(129) to compare the ratio of mCpG to CpG at PU, DM, and PU ESC Super Enhancer subregions in XEN and TS cells (figure 4.3A). This data is from the same publication referred to in figure 4.1G. In both XEN and TS cells PU appear CpG hypomethylated relative to DM and PM. However, the XEN and TS cells appear

hypermethylated at the PU relative to ESCs and EpiSCs. In both XEN and TS the DM and PM show similar distributions of mCpG/CpG. Therefore, a subset of PU may drive transcription in XEN and TS cells, whereas the DM only do so in ESCs.

Next I compared the CpG methylation between different extra-embryonic tissues *in vivo* using unpublished data shared by Dr Peter Rugg-Gunn (figure 4.3B). Post-implantation the TE and PrE respectively give rise to extra-embryonic ectoderm (ExE)(271) and visceral endoderm (VE). Dr Rugg-Gunn's group performed bisulphite-seq on dissected TE from E3.5 and E4.5 blastocysts and ExE and VE from E5.5 and E6.5 embryos. In all cell types at all embryonic days, PU appear hypomethylated relative to DM and PM. DM and PM gradually acquire CpG methylation in each cell type over time. Taken together, PU compared to DM and PM appear CpG hypomethylated in both pluripotent and extra-embryonic cells both *in vitro* and *in vivo* (E3.5-E6.5).

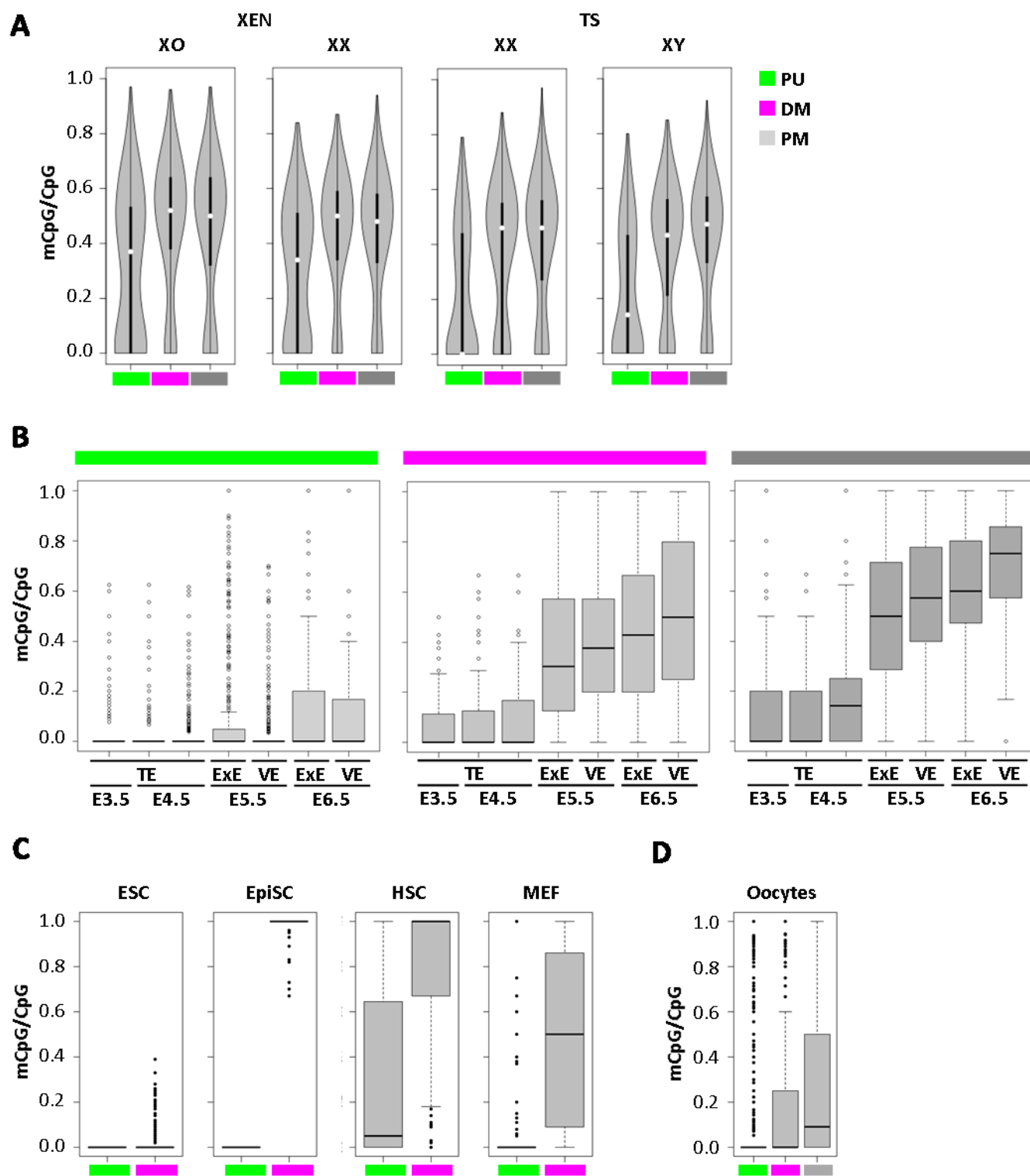
If PU lack CpG methylation in early pluripotent and extra-embryonic cells, what about more differentiated cell types? To test this I compared bisulphite-seq data generated from adult Haematopoietic Stem Cells (HSCs)(272) and Mouse Embryonic Fibroblasts (MEFs) derived from E12.5-E13.5 embryos (273,274) with ESCs and EpiSCs (figure 4.3C). Drs Alice and Luc Jouneau processed this data. Although both PU and DM display great variability, in both HSCs and MEFs PU compared to DM appear CpG hypomethylated. This suggests that at least a subset of PU may maintain activity in more differentiated cell types, as well as pluripotent and extra-embryonic tissues.

At birth, primary oocytes contain minimal CpG methylation. DNA methylation occurs progressively during oocyte growth, with the onset of DNA methylation coinciding with the transition from primary to secondary oocyte, and is completed by the time oocytes attain a diameter > 70 μ M. Remarkably, although most CGIs appear unmethylated in blastocysts, those that acquire methylation during oocyte maturation become CpG methylated in blastocysts.(275) Given that upon fertilisation the embryo undergoes global demethylation, this suggests that the epigenetic profile of pre-implantation blastocysts begins establishing

prior to fertilisation. To determine whether the oocyte foreshadows the epigenetic fates of Super Enhancelets, I examined the CpG methylation profiles of PU, DM and PM in oocytes using unpublished bisulphite-seq data provided by Dr Peter Rugg-Gunn (figure 4.3D). In oocytes CpG methylation of these subregions recapitulates the familiar PU<DM<PM hierarchy (Means: PU=0.14, DM=0.21. $P\text{-value}=2.781\times 10^{-5}$). Indeed, correlation testing (figure 4.3E) revealed that in each ESC Super Enhancer subregion, the oocytes and E3.5 ICM exhibit the strongest correlation scores (Spearman correlation co-efficient (ρ): PU - $\rho = 0.59$; DM - $\rho = 0.58$; PM - $\rho = 0.44$). Interestingly, the Spearman ρ values of the PU always positively correlated, whereas the DM Spearman ρ values show negative correlations. This evidences that the PU compared to the DM contain more similar CpG methylation profiles in all cell types at all time points. Density plotting the Spearman ρ values (with duplicates and same-to-same comparisons removed) confirms that the DM compared to the PU generate lower Spearman ρ values (Welch Two Sample t-test derived $P=0.001$) (figure 4.3F). These negative correlations between the E3.5-4.5 and E5.5-6.5 cells likely drive the clean segregation of pre-implantation and post-implantation derived cells observed in the clustering. This further supports the idea that the maturation of pluripotency involves the epigenetic repression of specifically the DM, whereas the PU maintain a permissive chromatin in multiple pluripotent, extra-embryonic, and more differentiated cell types.

4.2.4 Persistently Unmethylated and Differentially Methylated ESC Super Enhancelets display distinct epigenetic profiles in ESCs

In the chapter 3.2.3 I reported that while PM subregions compared to the unmethylated Super Enhancelets do contain greater levels of 5-mC, they show a much greater increase in 5-hmC and to a lesser extent 5-fC (figure 3.3C). Furthermore the Super Enhancelets display a bimodal distribution of 5-hmC implying two subpopulations. To investigate whether PU and DM Super Enhancelets contain differing 5-hmC and 5-fC I returned to the same data sets.(276,277) The Super Enhancelets show no difference in 5-mC content, however compared to PU the DM contain higher levels of 5-hmC and 5-fC (figure 4.4A).



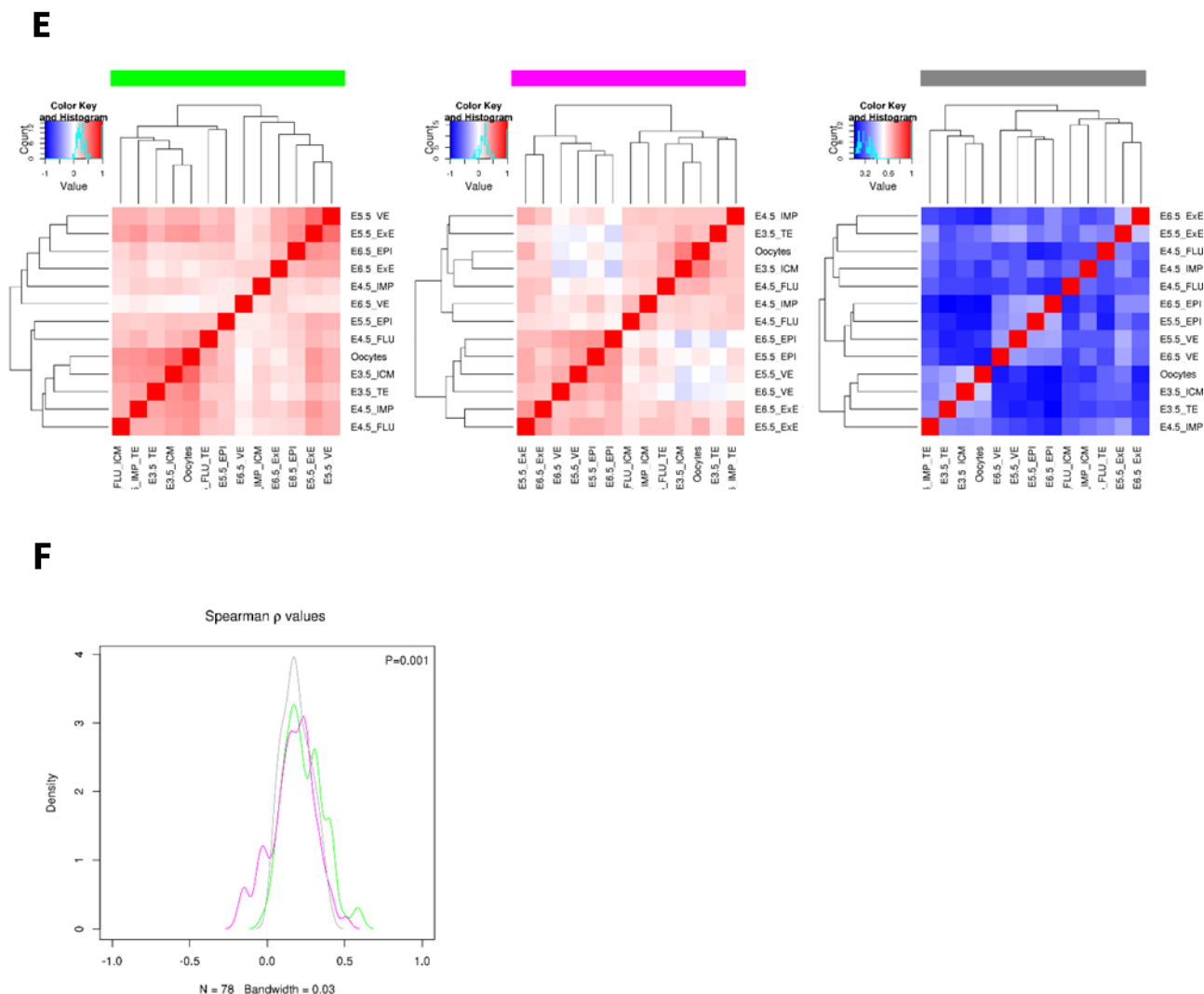


Figure 4.3 ESC Super Enhancelets display distinct CpG methylation profiles in non-pluripotent cell types.

A. Violin plots of MeDIP-seq data showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions, in XEN and TS cell lines. **B.** Boxplots of whole genome bisulphite-seq data showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in *in vivo* extra-embryonic tissue. (Unpublished data from Dr Peter Rugg-Gunn's group.) **C.** Boxplots of whole genome bisulphite-seq data showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in pluripotent (ESC and EpiSC) and differentiated cell lines (HSC and MEF). **D.** Boxplots of whole genome bisulphite-seq data showing the proportion of methylated CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in oocytes. (Unpublished data from Dr Peter Rugg-Gunn's group.) **E.** Heatmaps of the Spearman correlation values (ρ) between whole genome bisulphite-seq data derived from oocytes, pluripotent, and extra-embryonic tissues at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions. **F.** Density plot showing the distribution of Spearman correlation values (ρ) between whole genome bisulphite-seq data derived from oocytes, pluripotent, and extra-embryonic tissues at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions.

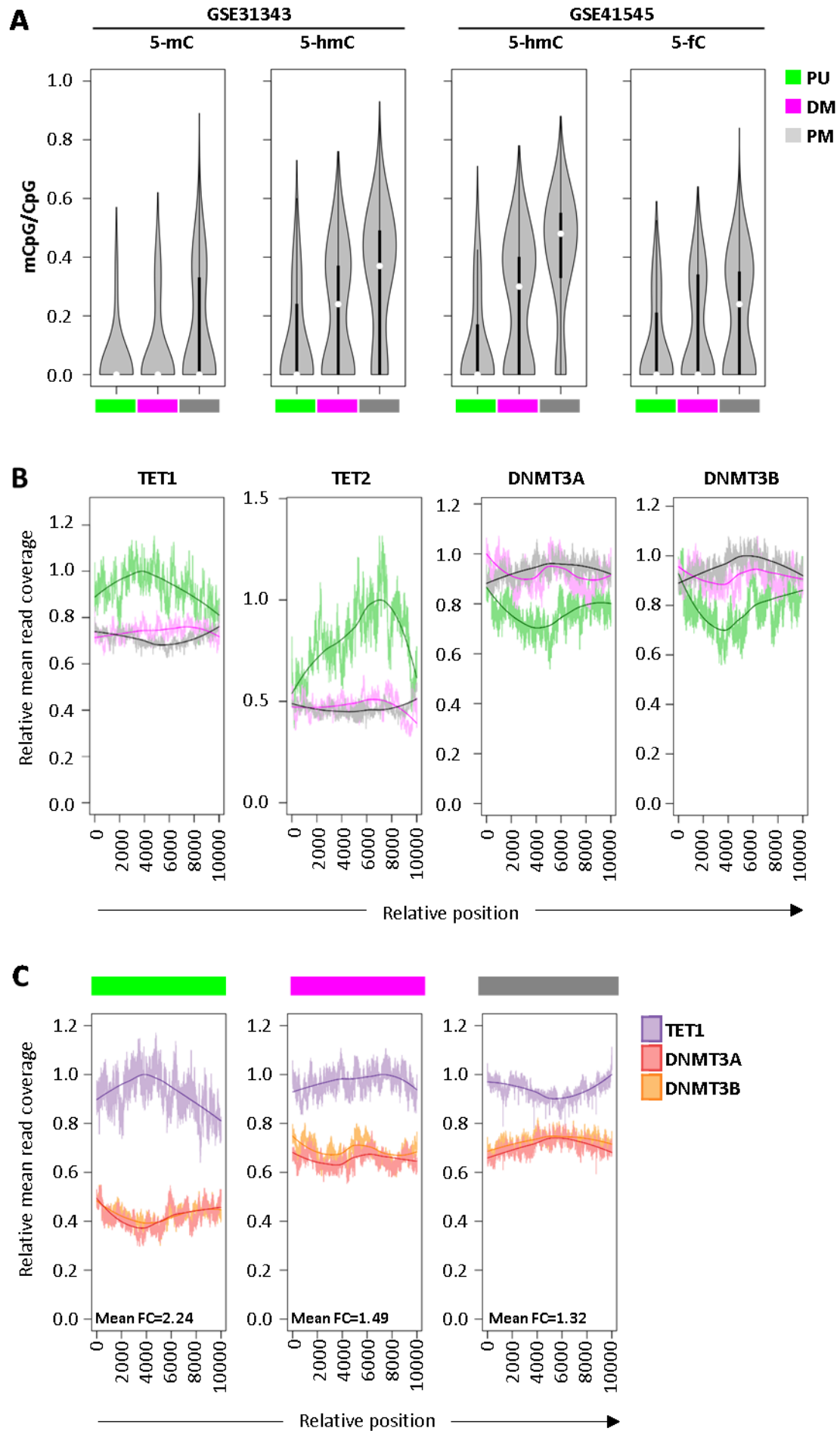


Figure 4.4 ESC Super Enhancelets display distinct CpG methylation profiles in ESCs.

A. Violin plots of DIP-seq showing the proportion of 5-mC, 5-hmC, and 5-fC CpGs at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in ESCs. **B.** Metagenes showing the mean ChIP-seq read coverage at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in ESCs. ChIP-seq bam files were normalised for sequencing depth and background (except in the case of the TET2 dataset where the input was not provided by the authors). The mean read coverage values at ESC Super Enhancers and Typical Enhancers was calculated using the `plotCoverage` function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value for either region set. The darker lines show the smoothed fit of the data generated by loess regression. B shows the metagenes by ChIP target and C by region set.

The TET family of enzymes catalyse the oxidation of 5-mC to 5-hmC and 5-fC.(60,63) Therefore, one might expect higher binding of both TET proteins and DNMTs to DM compared to PU. To check the protein binding density of these two groups of enzymes, I downloaded, processed, and calculated the relative mean sequencing read coverage of publicly available ChIP-seq data for TET1(258,278), TET2(279), DNMT3A(280) and DNMT3B(280) in ESCs grown in serum+LIF culture conditions (figure 4.4B).¹⁴ The authors of the TET2 ChIP-seq did not provide a control ChIP, thus I was unable to correct for mapability biases and the calculated read coverage will reflect chromatin accessibility to an unknown extent. Unexpectedly, the highest density of both TET proteins occurs at PU. DM and PM show roughly equivalent read coverage. Conversely, the DNMTs bind most densely to the DM and PM as predicted.

The metagenes in figure 4.4B show the difference in enrichment of a given protein between each subregion, but don't show the relative levels of each protein at single subregion. For example, the plots shown in figure 4.4B show us there are higher levels of TET1 at PU, compared to DM and PM in, but they don't inform us how much DNMT3A and DNMT3B is present relative to TET1 at PU. To investigate whether the ratio between TET and DNMTs may explain the counter-intuitive TET binding profiles, I calculated the relative mean read coverage of TET1, DNMT3A and DNMT3B at each of the PU, DM and PM (figure 4.4C). I

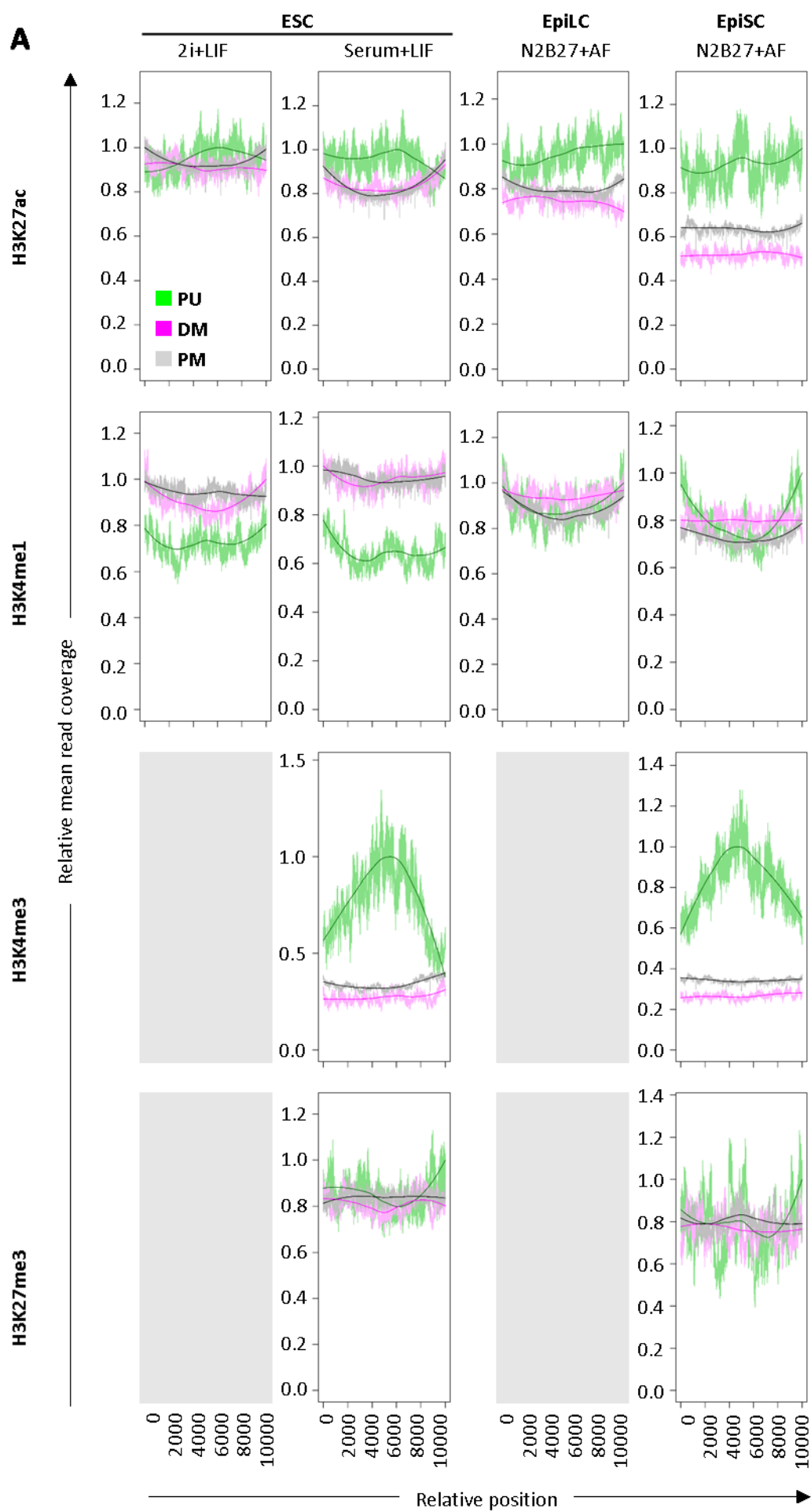
¹⁴ Methods: chapters 2.2.2.4.1, 2.2.2.4.1.3, 2.2.2.4.2

found a clear difference in the ratio of TET1 to the DNMTs at the PU versus the DM and PM. At each subregion, of the three proteins analysed TET1 bound most densely. The DM and PM show relative higher and equivalent levels of the DNMTs compared to the PU (mean fold-change between TET1 and DNMT3B coverage: PU = 2.24, DM = 1.49). This points to cycling of DNA methylation at the DM (and to some extent at the PM) compared to PU, which could explain the higher 5-hmC at the DM. Additionally, this could reflect a much briefer TET1-DNA interaction at DM compared to PU.

4.2.5 Persistently Unmethylated and Differentially Methylated enhancer units display different histone modification profiles during pluripotency maturation

In addition to DNA methylation, cells mediate the epigenetic regulation of gene expression via histone modifications. Histone modifications serve diverse purposes, for example: H3K27ac deposition associates with active promoters and enhancers;(251) H3K4me1 associates with both active and inactive enhancers(250,251,281); H3K4me3 in tandem with the repressive H3K27me3 associates with poised regulatory regions.(127,128) Given that the DNA methylation profiles of PU and DM Super Enhancelets differ, I postulated that they may bear differing patterns histone modifications. To test this hypothesis I downloaded, processed and calculated the relative mean sequencing read coverage of publicly available ChIP-seq experiments targeting the aforementioned histone modifications in ESCs grown in 2i+LIF(34) and serum+LIF,(251,282) as well as in EpiLCs(34) and EpiSCs.(283) At the time of this work only H3K27ac and H3K4me1 ChIP-seq data was publicly available in ESCs (2i+LIF) and EpiLCs. Figure 4.5A contains the resulting metagenes and figure 4.5B shows the mean fold-change in read coverage between the PU and DM for each histone modification in each cell type.

I reported in chapter 3 that in ESCs grown in serum+LIF, Super Enhancelets (unmethylated) and interstitial (methylated) regions of ESC Super Enhancers contain comparable levels of H3K27ac (see figure 3.3D). This holds true upon segregating PU and DM in both ESCs grown in 2i+LIF (fold change = 1.04) and serum+LIF (fold-change =1.15). The PU compared DM



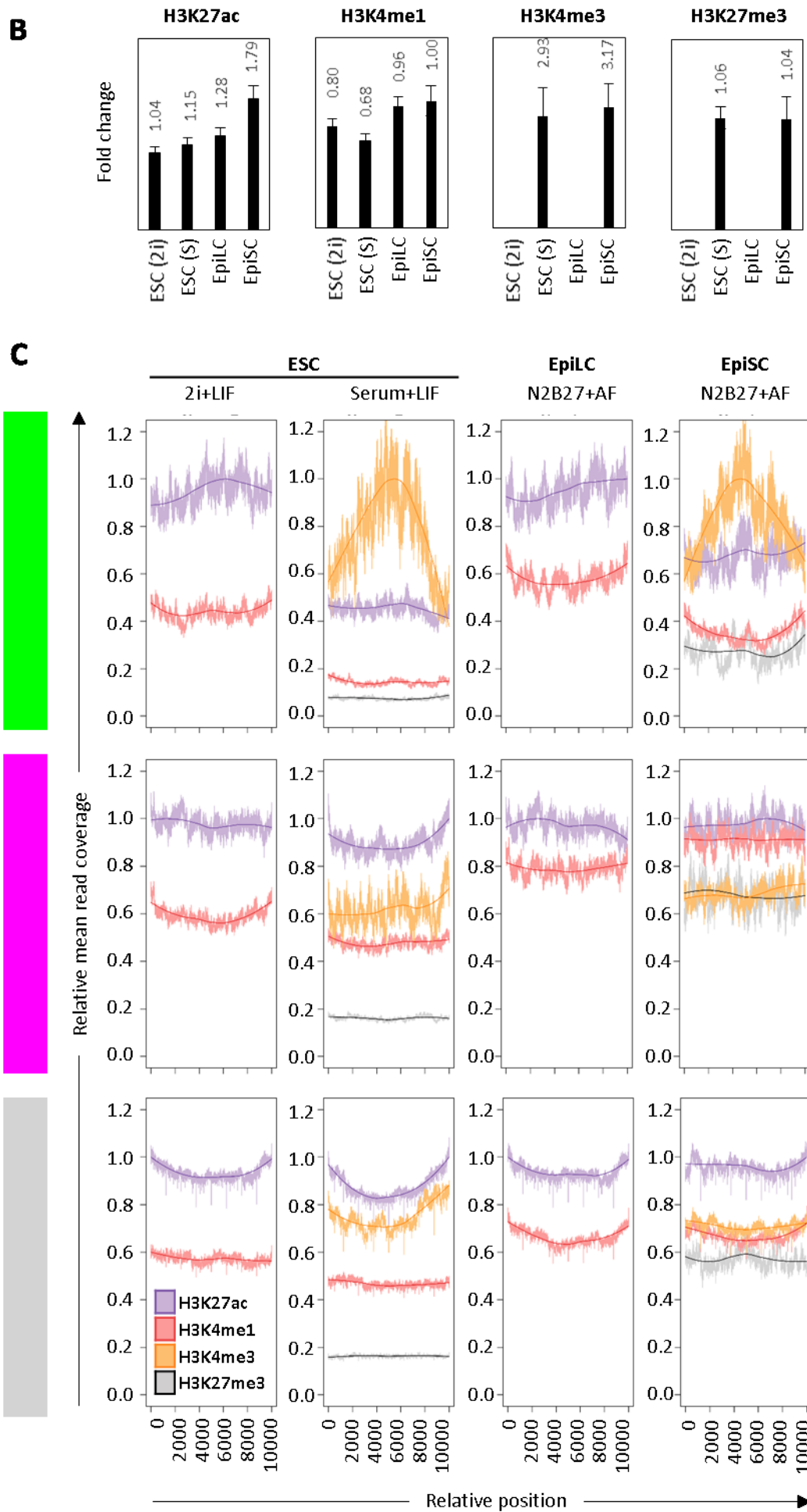


Figure 4.5 ESC Super Enhancers harbour different histone modification profiles in pluripotent stem cells.

Metagenes showing the mean ChIP-seq read coverage at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in ESCs grown in 2i+LIF and serum+LIF, EpiLCs, and EpiSCs. ChIP-seq bam files were normalised for sequencing depth and background. The mean read coverage values at each region set was calculated using the plotCoverage function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value. The darker lines show the smoothed fit of the data generated by loess regression. A plots the meta-genes by ChIP target and C by region set. B shows the fold-change in read coverage between PU and DM Super Enhancers in each cell type.

show a progressive increase in relative H3K27ac content in EpiLCs (fold-change = 1.28) and EpiSCs (fold-change = 1.79). This either means that H3K27ac content increases at PU or decreases at DM. H3K4me1 exhibits the opposite pattern. PU and DM show the greatest difference in ESCs (fold changes: 2i+LIF = 0.80; serum+LIF = 0.68), and the gap closes in EpiLCs (fold change = 0.96) and EpiSCs (fold change = 1.00). Surprisingly, given its canonical role as a mark of active promoters, PU contain 3-fold higher levels of H3K4me3 compared to DM in both ESC (fold change = 2.93) and EpiSC (fold change = 3.17). Several studies report the mutual exclusion of H3K4me3 and DNMTs.(284–287) King *et al.*, presented findings showing that H3K4me3 forms independently of DNA methylation.(288) Therefore, the high H3K4me3 at PU compared to DM could cause the lower relative levels of DNMT3A and DNMT3B found at PU (figure 4.4C). As H3K27me3 associates with repressed chromatin, it was unlikely that these active regulatory regions would contain biologically meaningful levels of H3K27me3. Indeed, of the 121,083 and 24,418 H3K27me3 ChIP-seq peaks called from the ESC and EpiSC data only 59 (0.05%) and 30 (0.12%) peaks overlapped with the ESC Super Enhancers. However, to discount the possibility of a minute enrichment at DM and PM compared to PU subregions I assessed the H3K27me3 metagenes. H3K27me3 appears evenly distributed across the three subregions in ESCs (fold change = 1.06) and EpiSCs (fold change = 1.04).

As with figure 4.4B, the metagenes in figure 4.5A show the difference in enrichment of a given histone modification between each subregion (i.e. there are equivalent levels of H3K27me3 at PU, DM and PM in ESCs), but don't show the relative levels of each histone modification at single subregion (i.e. how much H3K27me3 is present relative to H3K27ac at PU in ESCs?). To examine the relative histone modification content at each subregion, I computed the relative mean coverage per subregion (figure 4.5C). As expected, H3K27me3 exhibits the lowest enrichment of the four histone modifications at all subregions and cell types, in contrast to H3K27ac which exhibits the highest in almost all subregions and cell types. To my further surprise, H3K4me3 shows much higher enrichment than H3K27ac (a definitive Super Enhancer mark) at the PU in both ESCs and EpiSCs, which correlates with the CpG hypomethylation of this subregion in both these cell types. In DM in ESCs H3K4me3 enrichment approximates that of H3K4me1. Conversely, in EpiSCs H3K4me3 levels drop to that of H3K27me3.

In summary, PU and DM ESC Super Enhancers demonstrate distinct histone modification profiles in both naïve and primed pluripotency. Although one of the defining features of Super Enhancers, H3K27ac appears slightly enriched at PU compared to DM in ESCs grown in serum+LIF, and this enrichment becomes more pronounced as pluripotency matures. Predictably, all three subregions contain minimal H3K27me3. The relatively higher levels of H3K4me3 at PU compared to DM may inhibit the binding of DNMTs, potentially explaining the lower DNA methylation at PU. The complete lack of H3K4me3 at DM in EpiSCs suggests a rapid active histone demethylation process.

4.2.6 Differentially Methylated relative to Persistently Unmethylated enhancer units lose transcription factor occupancy during pluripotency maturation

In chapter 3 I reported that pluripotency regulatory proteins bind at higher density to unmethylated enhancer units (i.e. PU and DM) compared to methylated interstitial subregions (i.e. PM) ESC Super Enhancer subregions (figure 3.3D). If DM decommission as pluripotency matures, one might expect a concurrent reduction in protein binding. To investigate whether

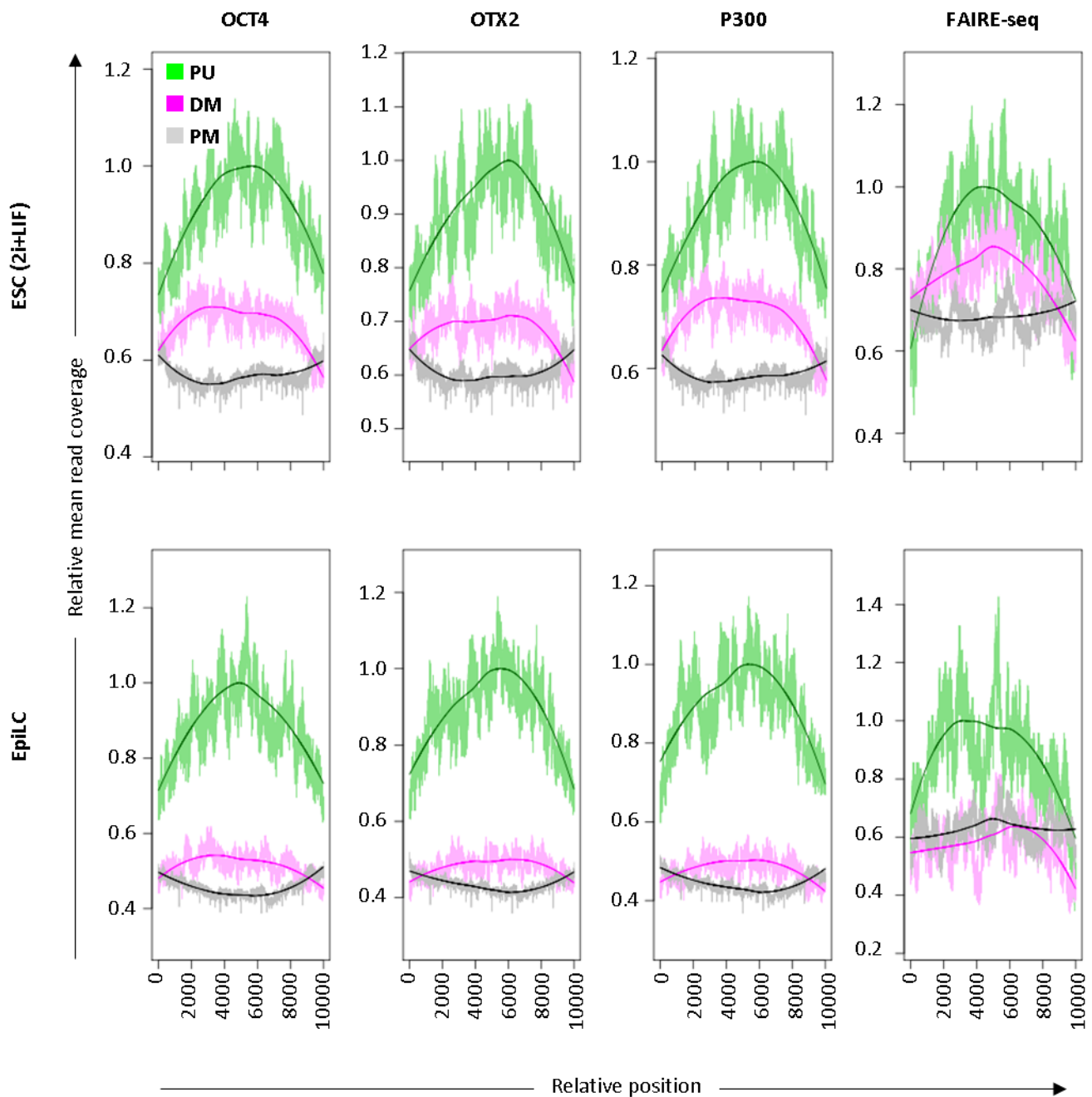


Figure 4.6 PU compared to DM ESC Super Enhancers bind higher densities of pluripotency transcription factors.

Metagenes showing the mean ChIP- and FAIRE-seq read coverage at PU (green), DM (magenta) and PM (grey) Super Enhancer subregions in ESCs (2i+LIF) and EpiLCs.(34) ChIP-seq bam files were normalised for sequencing depth and background. The mean read coverage values at each region set was calculated using the `plotCoverage` function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value. The darker lines show the smoothed fit of the data generated by loess regression.

this occurs I assessed the relative mean sequencing read coverage of key ESC and EpiSC transcription factors at PU, DM and PM in ESC (2i+LIF) and EpiLCs (figure 4.6).⁽³⁴⁾ OCT4 acts as a pioneer factor in both naïve and primed pluripotent stem cells, and localises to active enhancers along with P300 and high H3K27ac deposition.^(289–292) OTX2 joins OCT4 in this role in EpiLCs and EpiSCs only.⁽³⁴⁾ P300 acetylates H3K27.⁽²⁹³⁾ In ESCs OCT4, OTX2 and P300 density at Super Enhancer subregions follows the PU>DM>PM pattern. Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-seq) isolates nucleosome-depleted chromatin (i.e. active regulatory regions).⁽²⁹⁴⁾ Although not as cleanly separated as the transcription factors, PU appear the most nucleosome-depleted after DM and then PM. The EpiLCs display starkly different patterns. For both the transcription factors and the FAIRE-seq the DM sequencing read densities reduce to similar levels of that of the PM. This suggests that only the PU retain binding of OCT4, OTX2 and P300, as well as an open chromatin state in EpiLCs. The reduction in EpiLCs at DM of OCT4, OTX2, and P300 binding, in conjunction with the chromatin condensation as indicated by the FAIRE-seq, implies a decrease in enhancer activity from these regions.

4.2.7 The selective retention of unmethylated Super Enhancelets defines two distinct classes of ESC Super Enhancers with differential activity in EpiSCs and epiblast *in vivo*

So far we have seen that a subset of ESC Super Enhancelets (DM) selectively acquire CpG methylation during pluripotency maturation, while the remaining Super Enhancelets (PU) maintain an unmethylated state in both cell types. In tandem, DM lose binding of pluripotency-related transcription factors and permissive histone modifications. Furthermore these two types of Super Enhancelet display distinct transcription factor binding and epigenetic profiles in ESCs. These results point to differential regulatory mechanisms active at these regions. Based on Super Enhancelet complement I divided Super Enhancers into two classes: class I – Super Enhancers that contain at least one PU subregion; class II – Super Enhancers that only contain DM subregions (figure 4.7A, appendix A1.1.2). I hypothesised that genes

associated with class I Super Enhancers may retain gene expression in EpiSCs, explaining the heterogeneity in differential expression between ESCs and EpiSCs that I observed in chapter 3 (figure 3.6C-E). To test this hypothesis I employed RNA-seq data from ESCs (both grown in 2i+LIF and serum+LIF)(134) and EpiSCs.(130) Dr Alice Jouneau's lab generated the EpiSC data, and she and Dr Luc Jouneau processed all of these datasets for the purposes of this experiment.¹⁵ I assessed the expression of class I and II Super Enhancer associated genes by comparing the $\log_{10}$ transformed Reads Per Kilobase of transcript per Million mapped reads (RPKM) and presented these results as violin plots (figure 4.7 B). Class I associated genes (green) show a minor and non-significant decrease in expression in EpiSCs (mean fold change = 0.79, $P = 0.07$). However, class II associated genes (pink) showed a dramatic decrease in expression in EpiSCs (mean fold change = 0.55, $P = 0.001$). To validate this result *in vivo*, I employed privately shared RNA-seq data for ICM and epiblast (E6.2 and E6.7) processed by Drs Jouneau (Bernado *et al.*, in submission). Given the close time points of the two epiblast samples, I performed a correlation test to compare their distributions. To determine whether to use the parametric Pearson or non-parametric Spearman correlation test I used the Shapiro-Wilk test for normality. The P-values for the distribution of $\log_{10}$ (RPKM) were 4.094×10^{-10} and 3.213×10^{-08} and therefore I rejected the null hypothesis that the data were normally distributed and used the non-parametric Spearman correlation test. The E6.2 and E6.7 epiblast samples returned a Spearman's ρ of 0.92 indicating high similarity between the two distributions. The $\log_{10}$ (RPKM) values of the two classes of ESC Super Enhancer associated genes in ICM and epiblast mimic the *in vitro* result (figure 4.7C) (Wilcoxon rank sum test with continuity correlation P-values: class I = 0.13, class II = 0.0001). These results indicate that only the class II ESC Super Enhancer associated genes downregulate as pluripotency matures.

Given that the class I and class II ESC Super Enhancer associated genes demonstrate different expression patterns in primed pluripotency, I proposed that they regulate distinct sets

¹⁵ Methods: chapter 2.2.2.4.1.5

of genes. To test this I performed a Gene Ontology (GO)(295,296) over-representation test on gene sets associated with class I and class II ESC Super Enhancers. The GO results indicate that the two classes enrich for different biological processes (table 4.1, figure 4.7C, appendices A.1.1.5 and A.1.1.6). Class I genes enrich for positive transcriptional regulation, key cell signalling pathways (cellular response to leukaemia inhibitory factor, regulation of MAPK cascade), and multiple differentiation processes (neural tube closure, endodermal cell fate specification, regulation of cell differentiation, response to retinoic acid). The Class II associated gene set returns fewer results, possibly due to class II containing fewer genes than class I, which may result in lesser statistical power to find enrichment (Class I = 129, Class II = 94). The Class II genes enrich for negative transcriptional regulation, epithelial differentiation processes, and generic chordate embryonic development (i.e. of the Chordata phylum, defined as animals possessing a notochord, essentially a spinal column). Only one parent term overlaps between the classes: 'transcription, DNA-templated (GO:0006351)'. These results suggest that Class I genes contribute to the priming of pluripotency, potentially even lineage commitment, and association with the MAPK pathway, which is a key mode of gene expression regulation in EpiSCs.(10,11) That the majority of Class II GO terms relate to negative transcriptional regulation may suggest that these genes suppress transcription, perhaps of genes required to drive development onwards.

To further characterise the genes associated with Class I and Class II ESC Super Enhancers I performed a competitive mean-rank gene set test¹⁶. This tests whether the genes in a given pathway tend toward differential expression compared to genes not in the pathway. The gene set tests requires two input parameters: a list of a statistics, and a means by which to index them. I used the negative $\log_{10}$ transformed P-values (corrected for multiple hypothesis testing) for differential mRNA expression between ESCs and EpiSCs, as the list of statistics, and for each pathway set the index to the genes included in Class I or Class II. Given the small number of genes in each class there was insufficient power to statistically calculate

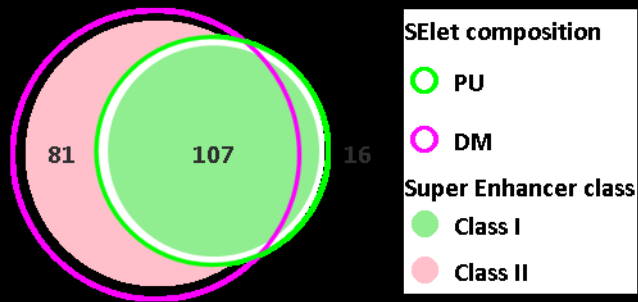
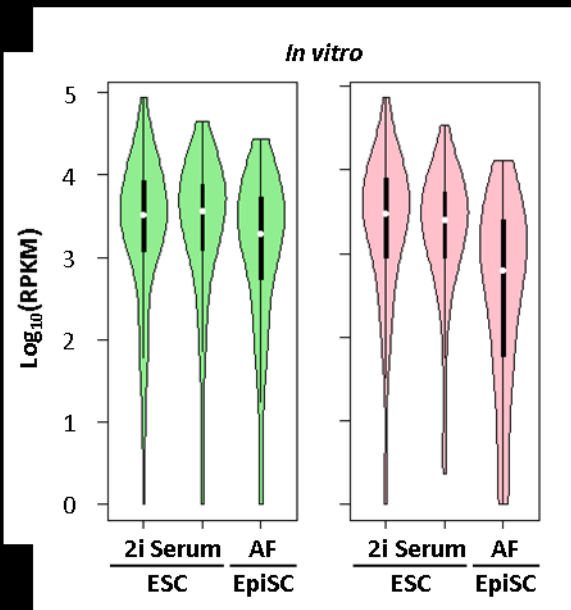
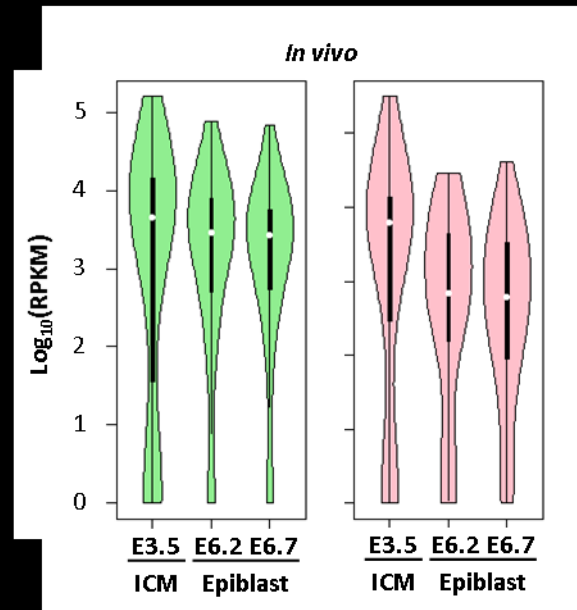
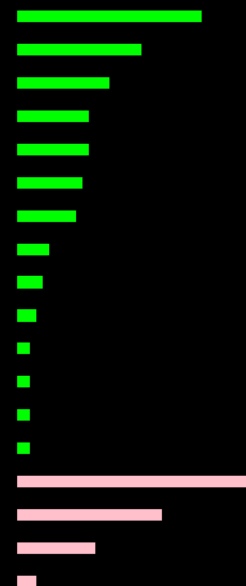
¹⁶ Methods: chapter 2.2.2.5.2

Table 4.1 Gene Ontology terms enriched in Class I and II ESC Super Enhancer associated gene sets.

The Gene Ontology parent terms enriched in Class I and II ESC Super Enhancer associated gene sets, with the count of enriched GO terms. GO terms are organised hierarchically. For example, 'GO:0004923 leukemia inhibitory factor receptor activity' belongs to the parent term 'GO:0048861 leukemia inhibitory factor signaling pathway', which in turn belongs to 'GO:0007167 enzyme linked receptor protein signaling pathway'. This table lists the parent terms enriched in Class I and II Super Enhancer associated genes along with the number of GO terms belonging to that parent term that show enrichment. The parent term highlighted in grey was found to be enriched in both the Class I and II gene sets.

Class	GO Parent Term	Count of GO Terms
I	positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	28
I	neural tube closure (GO:0001843)	19
I	endodermal cell fate specification (GO:0001714)	14
I	cellular response to leukemia inhibitory factor (GO:1990830)	11
I	regulation of MAPK cascade (GO:0043408)	11
I	transcription, DNA-templated (GO:0006351)	10
I	negative regulation of intracellular signal transduction (GO:1902532)	9
I	negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	5
I	negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	4
I	somatic stem cell population maintenance (GO:0035019)	3
I	regulation of cell differentiation (GO:0045595)	2
I	regulation of multicellular organismal development (GO:2000026)	2
I	response to retinoic acid (GO:0032526)	2
I	transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	2
II	negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	35
II	transcription, DNA-templated (GO:0006351)	22
II	epithelial tube morphogenesis (GO:0060562)	12
II	chordate embryonic development (GO:0043009)	3

enrichment. Therefore I outputted ranks of the P-values for enrichments, rather than the P-values. This allows the identification of trends in gene expression. Appendix A.1.1.5 details the pathways with a rank of 0.5 or lower (the maximum rank being 1). Figure 4.7E shows the density of $\log_2$ transformed fold-changes in expression for all genes in a selection of biologically relevant pathways. These pathways return a rank of 0.5 or lower for one or both classes of ESC Super Enhancer associated gene, and contain at least members of either class. The mean (dotted lines) and standard deviations (coloured rectangles) in expression of the Class I and II genes contained within that pathway are overlaid. The gene set test returned pathways involved in the biological processes identified as enriched by the GO, for example MAPK signalling (Ras signalling pathway) and neuronal lineation. The plots show fewer Class II compared to Class I genes are contained within these pathways. Notably, Class I genes



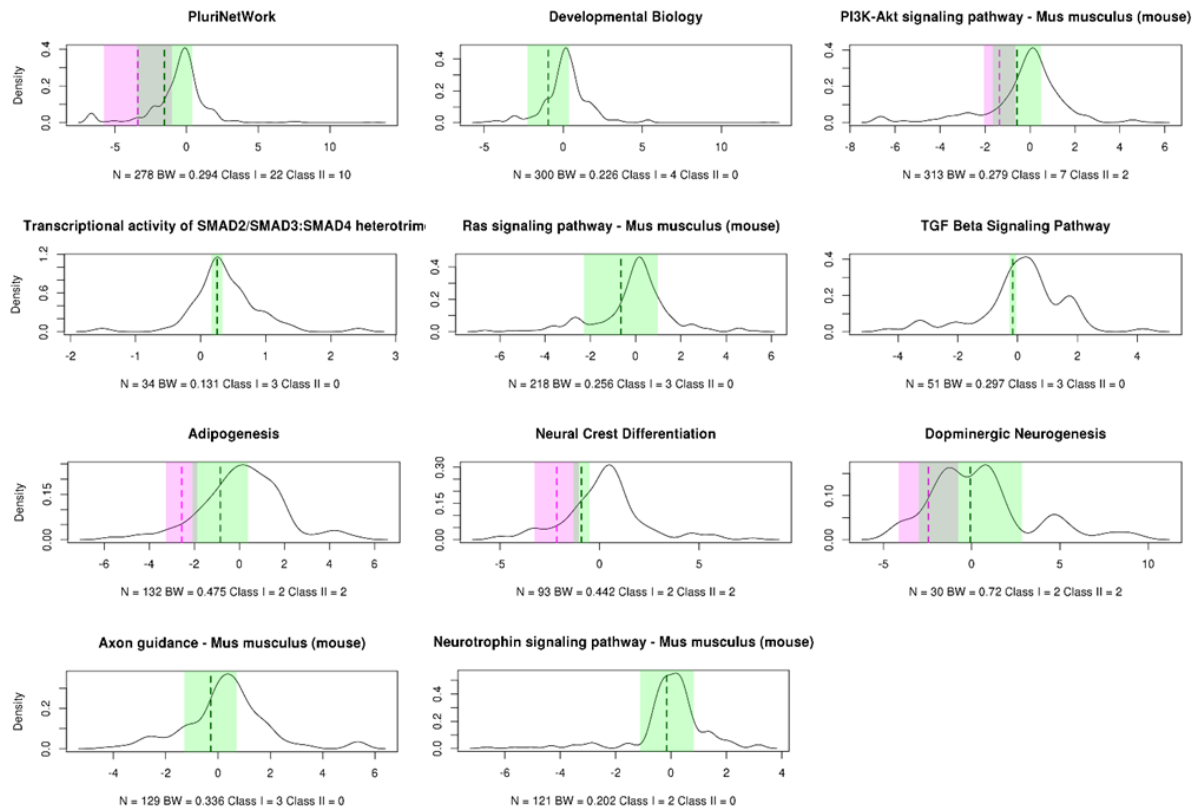
E

Figure 4.7 Only genes associated with Class II ESC Super Enhancers tend to reduce transcription as pluripotency matures.

A. Venn diagram showing the number of Super Enhancers composed of PU and/or DM Super Enhancelets. **B.** Violin plots of RNA-seq data show the $\log_{10}(\text{RPKM})$ of Class I (light green) and II (pink) Super Enhancer associated genes in ESCs grown in 2i+LIF and serum+LIF, and EpiSCs grown in N2B27 with Activin and bFGF (AF). **C.** Violin plots of RNA-seq data show the $\log_{10}(\text{RPKM})$ of Class I (light green) and II (pink) Super Enhancer associated genes in ICM and epiblast tissue. **D.** Barplot showing the number of Gene Ontology (GO) terms under each parent terms enriched for Class I (light green) and II (pink) Super Enhancer associated genes. **E.** Density plots showing the $\log_2(\text{fold-change})$ values of genes in pathways statistically significantly differentially expressed according to a competitive mean-rank gene set test. The dotted line indicates the mean expression of the Class I (green) and II (pink) genes within that pathway and the coloured rectangle represents the standard deviation.

show maintained transcription in these pathways, whereas Class II genes appear comparatively downregulated.

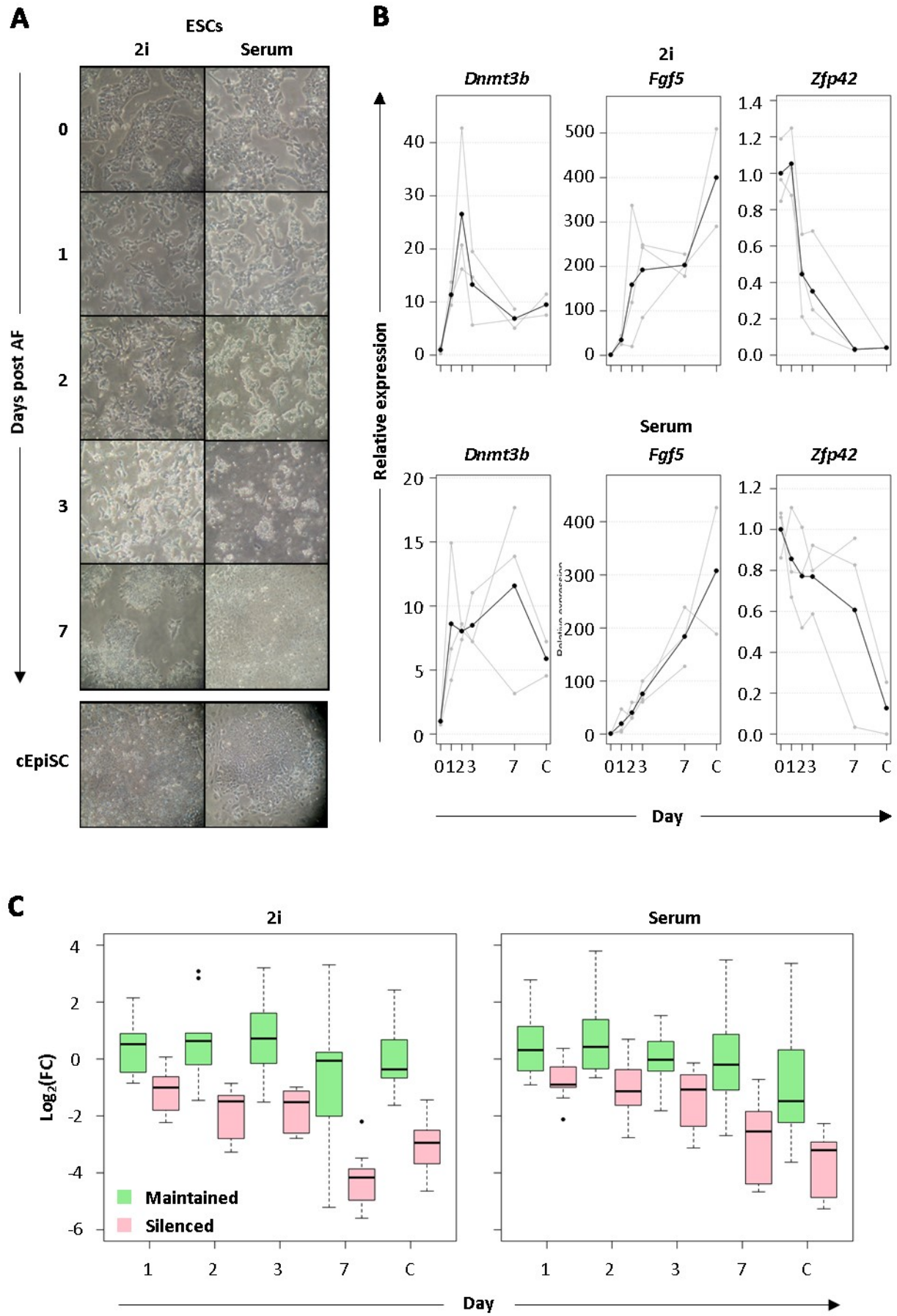
In summary, ESC Super Enhancers can be divided into two classes based on the differential CpG methylation of their Super Enhancelets in ESCs and EpiSCs. Class I ESC Super Enhancers contain PU Super Enhancelets, whereas Class II ESC Super Enhancers contain only DM Super Enhancelets. Class I ESC Super Enhancer associated genes maintain transcription in EpiSCs. In contrast, Class II ESC Super Enhancer associated genes reduce transcription in EpiSCs. Classification of these genes by GO and gene set testing reveals that Class I and II genes drive different biological processes. Henceforth, Class I and II Super Enhancers will be referred to as Maintained (I) and Silenced (II).

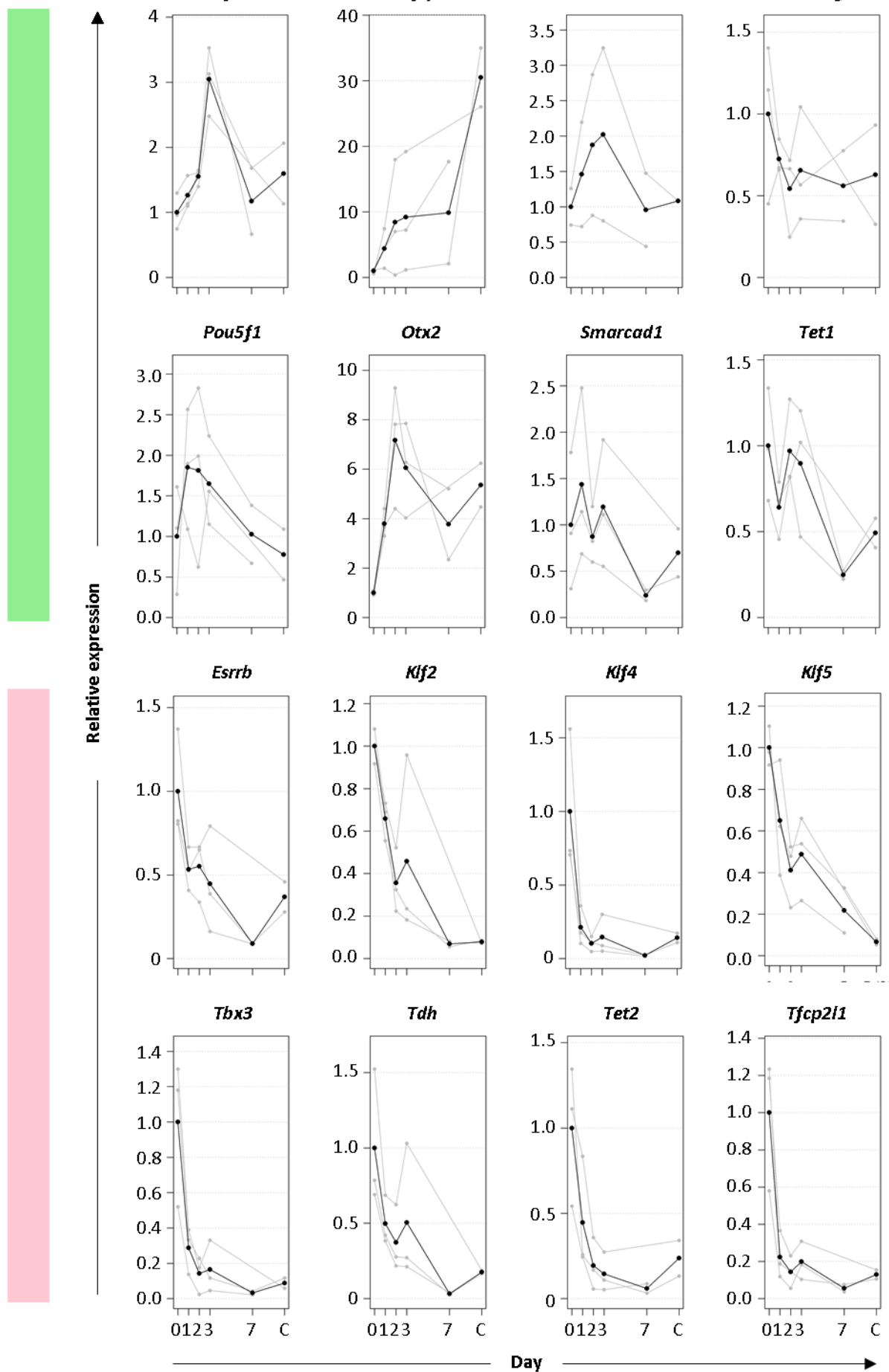
4.2.8 Class I and II super-enhancer associated genes show differential expression patterns upon ESC-to-EpiSC conversion in vitro

The experiments described in chapter 4.2.7 pertain to parental ESC and EpiSC cell lines propagated in culture for multiple passages. To determine whether the transcriptional fates of Maintained and Silenced ESC Super Enhancer associated genes truly diverge in tandem with pluripotency maturation, we performed RT-qPCR¹⁷ analysis of candidate genes during the conversion of ESCs to EpiSCs¹⁸, extracting RNA at days 0, 1, 2, 3 and 7 after the addition of Activin and fibroblast growth factor (FGF).^(10,11) The conversion was conducted from both 2i+LIF and serum+LIF starting conditions. Figure 4.8A shows photographs of the cells during the course of the conversion. The cells clearly take on a broader, epithelial-like morphology characteristic of EpiSCs. In these experiments, and as expected, we observed a characteristic loss of the naïve pluripotency marker *Zfp42* (REX1) in parallel with the induction of primed pluripotency factors *Fgf5* and *Dnmt3b* (figure 4.8B). Strikingly, all Silenced ESC SE associated genes analysed begin to lose expression 1 day post-induction (figure 4.8C, figure 4.8D lower panel, figure 4.8E lower panel). The Maintained ESC SE associated genes however retain a

¹⁷ Methods: chapter 2.2.1.4

¹⁸ Methods: chapter 2.2.1.3



D

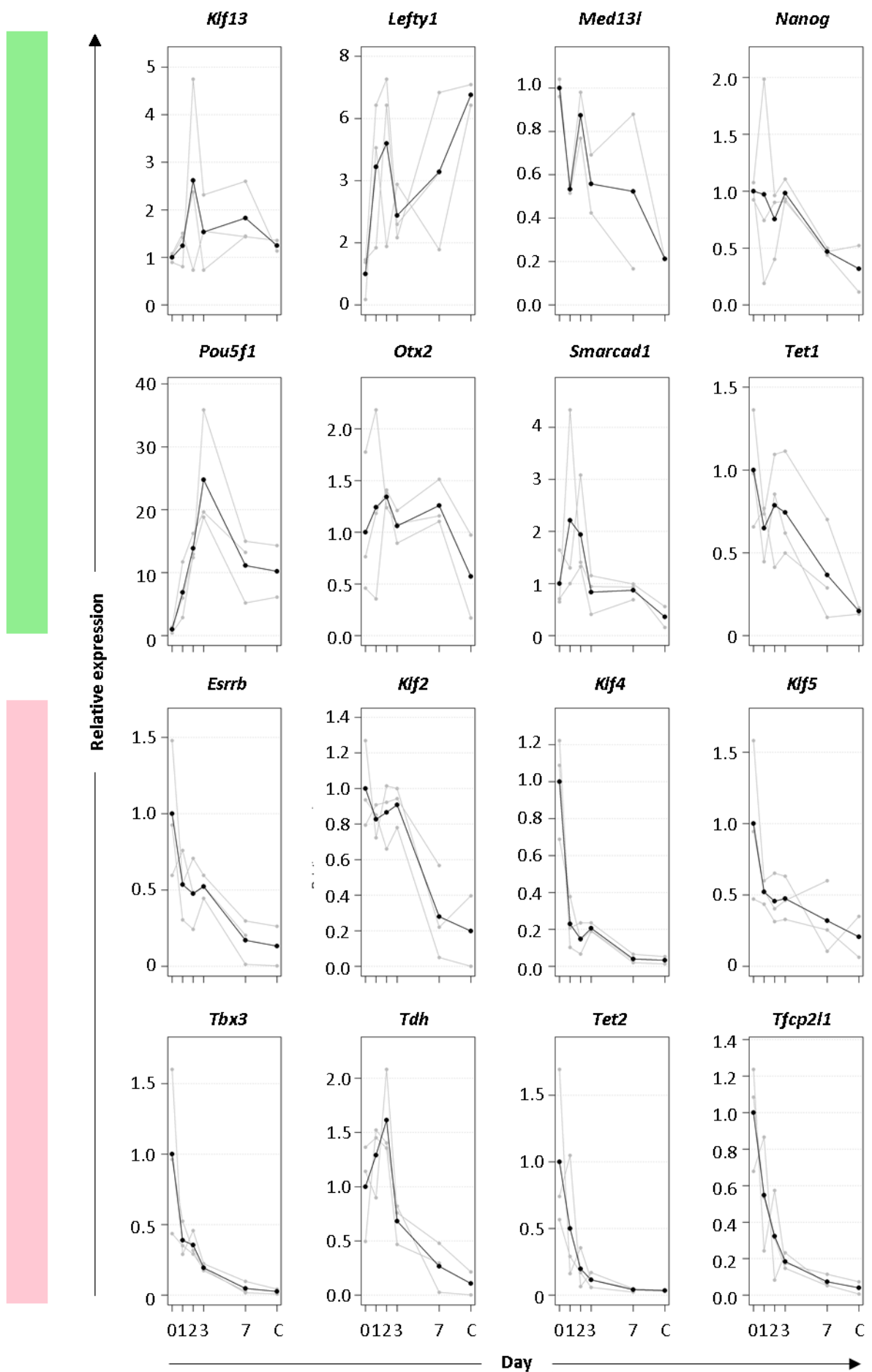
E

Figure 4.8 Class II ESC Super Enhancer associated genes gradually reduce transcription during ESC-to-EpiSC conversion, whereas Class I maintain expression.

A. Photographs of the cells during ESC-to-EpiSC conversion and after stable several passages (cEpiSC). **B.** Line graphs of RT-qPCR data showing the relative expression of the phenotypic control genes *Dnmt3b*, *Fgf5*, and *Zfp42* during ESC-to-EpiSC conversion. **C.** Boxplots showing the $\log_2(\text{fold-change})$ in expression of Class I (light green) and II (pink) ESC Super Enhancer associated genes. **D.** Line graphs of RT-qPCR data showing the relative expression of the Class I (light green) and II (pink) ESC Super Enhancer associated genes during ESC-to-EpiSC conversion, starting from ESCs grown in 2i+LIF. **E.** Line graphs of RT-qPCR data showing the relative expression of the Class I (light green) and II (pink) ESC Super Enhancer associated genes during ESC-to-EpiSC conversion, starting from ESCs grown in serum+LIF.

median $\log_2$ fold change of approximately 0 at all time-points except day 7, where the range in expression becomes highly variably and the median dips to approximately $\log_2 -1$ (figure 4.8C, figure 4.8D upper panel, figure 4.8E upper panel). This holds true for *Pou5f1* (OCT4) and *Otx2* whose expression is essential for the maintenance of EpiSCs(290) as well as *Tet1*, which remained highly expressed in the earliest steps of ESC-to-EpiSCs as previously reported (figure 4.8D upper panel, figure 4.8E upper panel).(130) In summary, the genes associated with Silenced ESC Super Enhancers increasingly lose expression over the course of ESC-to-EpiSC conversion. In contrast, the genes associated with Maintained ESC Super Enhancers either preserve or increase in expression.

4.3 Conclusion

The experiments described in Chapter 4: Results I demonstrate that total CpG methylation at ESC Super Enhancers increases as pluripotency matures. In tandem, their associated genes slightly but consistently increase CpG methylation at their promoters, and broadly reduce transcription. Initially these observations led us to conclude that ESC Super Enhancers decommission in EpiSCs, however closer examination revealed that some of the constituent enhancers, here termed Super Enhancelets, retain their CpG unmethylated state in epiblast both *in vitro* (figure 4.1) and *in vivo* (figure 4.2). We labelled these two types of Super Enhancelets Persistently Unmethymated (PU) and Differentially Methylated (DM). The

remaining interstitial CpG methylated subregions we called Persistently Methylated (PM). Although PU and DM ESC Super Enhancelets harbour similarly minimal levels of CpG methylation in ESCs, a number of differences emerged during these analyses. PU compared to DM contain a greater normalised CpG fraction (figure 4.1D). Relative to DM, PU appear CpG hypomethylated in extra-embryonic, more differentiated cell lines, and even in oocytes (figure 4.2). In fact, the CpG methylation levels of PU, DM and PM in oocytes predicts that of the E3.5 ICM (figure 4.3E). While PU and DM contain similar levels of 5-mC, DM show enrichment for the oxidation products 5-hmC and 5-fC (figure 4.4A). The increased enrichment of TET1 to DNMT3A/3B binding at PU compared to DM may contribute to this (figure 4.4C). In turn, the relatively higher density of H3K4me3 at PU compared to DM may inhibit the binding of the DNMTs (figure 4.5C). Nucleosome-depletion (i.e. regulatory region activity), as measured by FAIRE-seq, mimics the change in protein binding. In ESCs nucleosome-depletion follows the PU>DM>PM pattern, but in EpiLCs DM and PM show equivalent levels implying the condensing of DM chromatin. We divided the ESC Super Enhancers into two classes based on their Super Enhancelet composition: Class I contain at least one PU, whereas Class II contain only DM (figure 4.7A). The genes associated with these two classes of ESC Super Enhancer show different patterns of gene expression in EpiSCs *in vitro* (figure 4.7B) and *in vivo* (figure 4.7C). Class I genes maintain their expression, whereas Class II genes become silenced. Henceforth we term these classes Maintained and Silenced. GO (figure 4.7C) and gene set testing (figure 4.7D) of these two classes revealed that they regulate distinct biological processes and pathways. Finally, tracking gene expression changes during ESC-to-EpiSC conversion showed that these changes occur in response to Activin and FGF exposure (figure 4.8C). In summary, we define two types of Super Enhancelet with discrete transcriptional and epigenetic profiles. Collectively, the results presented here suggest that the DM decommitment in EpiSCs – therefore perhaps truly representing naïve pluripotency – whereas the PU maintain activity.

5 RESULTS III

ESRRB REGULATES THE ACTIVITY OF DM ESC SUPER ENHANCELETS

5.1 Introduction

In Chapter 4: Results we established that total CpG methylation at ESC Super Enhancers increases during the transition from naïve to primed pluripotency. At the same time, the genes associated with ESC Super Enhancers reduce transcription. These two observations suggest that ESC Super Enhancers decommission in EpiSCs. However, closer examination of constituent enhancers across ESC Super Enhancers, here termed Super Enhancelets, revealed that a fraction of them selectively retain their CpG unmethylated state in epiblast both *in vitro* (figure 4.1) and *in vivo* (figure 4.2) whilst other constituents and interstitial sub-regions become hypermethylated. We call these two types of Super Enhancelets Persistently Unmethylated (PU) and Differentially Methylated (DM), and the remaining interstitial CpG methylated subregions Persistently Methylated (PM). Based on their Super Enhancelet composition, we divided the ESC Super Enhancers into two classes: Class I contain at least one PU, whereas Class II contain only DM (figure 4.7A). In EpiSCs *in vitro* (figure 4.7B) and *in vivo* (figure 4.7C), Class I Super Enhancer-associated genes maintain their expression, whereas Class II genes become silenced. We therefore term these two classes of ESC Super Enhancers Maintained (I) and Silenced (II). In this chapter, I endeavour to identify members of the ESC transcriptional network driving the differing behaviours of these two classes of enhancers in naïve ESCs.

5.2 Results

5.2.1 shRNA screen identifies Silenced ESC Super Enhancer associated genes most sensitive to the loss of naïve pluripotency factors

Maintained and Silenced ESC Super Enhancers exhibit different epigenetic (i.e. DNA methylation and histone modification) profiles in ESCs and EpiSCs (see Chapter 4). Concurrently, their associated genes display different patterns of regulation. To identify members of the ESC transcriptome that exclusively regulate Maintained or Silenced ESC Super Enhancers in naïve ESCs, I employed a publicly available shRNA screen to compare the expression of their affiliated gene sets.⁽²⁹⁷⁾ In this study, Nishiyama *et al.* inhibited the expression of 100 transcription factors by co-transfecting ESCs with shRNAs cloned into a pPyCAG-EGFP-IP expression vector that carries puromycin gene resistance gene. Transfected cells were accordingly cultured under puromycin selection, ensuring the sole growth of cells that had successfully taken up the transfection vectors. Here, most of the candidate factors were targeted using two independent shRNAs, and the resulting cells treated as biological replicates. In three instances (*Nanog*, *Pou5f1*, *Nr0b1*) transfected cells were split, and the samples treated as technical replicates. After 72 hours of puromycin selection, cells were harvested, total RNA extracted and then applied to microarrays to evaluate the effect of the shRNA inhibition on the ESC transcriptome (see figure 5.1A a schematic of their experimental design). As a control, they treated ESCs with transfection reagent without DNA, and in the absence of puromycin selection. I downloaded from GEO (GSE26520) the probe intensity values for each array, log₂ transformed and quantile normalised them, applied a linear regression model and quantified differential expression through moderated t-statistics derived from that model between each shRNA treated ESC microarray and the control shRNA microarray.¹⁹

¹⁹ Methods: Chapter 2.2.2.3

To measure the knock down efficiencies of the different shRNAs used, Nishiyama *et al.* first examined the expression of their respective target gene according to microarray and RT-qPCR. To check that my method of microarray processing replicated the results published by Nishiyama *et al.*, I tested the Pearson correlation co-efficients between our calculated differential expression of the target shRNAs. Figure 5.1B and appendix A.1.1.6 show the $\log_2(\text{fold-change})$ and $-10 \times \log_{10}(\text{adjusted P-values})$ pertaining to my calculated target shRNA knock downs. ESCs containing shRNAs targeting *Dmrt2*, *Ncl*, *Batf3*, *Gabpa*, *4932441K18Rik*, *Foxd3*, *Jarid2*, *Eomes*, and *Sox1* showed non-significant differential expression compared to the control cells (i.e. $\log_2(\text{fold-change}) > -1$ and an adjusted P-value > 0.05). I discounted these arrays from further analysis. ESCs containing shRNAs targeting *Fem1b*, *Mkrn1*, *Foxp1*, *Smad7*, *Zscan10*, *Etv1*, *Sall1*, *T*, *Zfp57*, *Dppa3*, *Sox15*, *Sfrs6*, *Tcea3*, *Pml*, *Tcfap4* showed a small, but significant decrease in expression (i.e. $-1 > \log_2(\text{fold-change}) > 0$, but a P-value < 0.05). I continued to include these arrays, but considered the results with caution. I observed a strong positive correlation ($\rho = 0.74$, P-value = 0) between my and Nishiyama *et al.*'s microarray derived values. Microarrays have long been criticised for lack of reproducibility and specificity relative to RT-qPCR. Uneven and non-specific hybridisation, seasonal fluctuations in ozone levels, and differences in background correction and normalisation practices can all impact probe intensity values.(298) Therefore, it is critical to evaluate findings by RT-qPCR. The shRNA knock down efficiencies from Nishiyama *et al.*'s microarray and RT-qPCR values show a weaker, but still positive correlation ($\rho = 0.51$, P-value = 4.29×10^{-08}). I observed a similar level of correlation between my processed microarray and Nishiyama *et al.*'s published RT-PCR values ($\rho = 0.46$, P-value = 1.12×10^{-06}). Given the strong similarity between Nishiyama *et al.*'s and my evaluation of the shRNA knock down efficiencies, I conclude that my method of microarray processing successfully reproduces Nishiyama *et al.*'s data.

Before looking at the differential impact of each shRNA on the expression of Maintained and Silenced Super Enhancer- associated genes, I first wished to examine the behaviour of all Super Enhancer associated genes. To do this I calculated and ranked the median t-statistic

for differential expression of all ESC Super Enhancer associated genes upon shRNA knock down of the panel of key ESC transcription factors relative to control samples (figure 5.1C and table 5.1). I used the t-statistic as it reflects both the directionality and significance of differential expression. Of the 100 shRNA targets tested (109 included replicates), only 22 exhibit a median $|t| > 2$, all of which reflect a downregulation in overall expression. These shRNA targets include key ESC transcription factors such as the naïve pluripotency drivers *Esrrb* and its co-factor *Ncoa3*, *Klf4* and *Tbx3* and the general pluripotency drivers *Pou5f1* (OCT4), *Sox2* and *Nanog*. This result echoes those reported in chapter 3.2.6 (figure 3.6B), where I reported that naïve and general pluripotency factors reduce in expression in tandem with ESC Super Enhancer associated genes.

Table 5.1 Super Enhancer associated genes for which the microarray platform used by Nishiyama et al. had no target probes.

Maintained	Silenced
<i>Fam134b</i>	<i>BC096441</i>
<i>Fam53a</i>	<i>Cenpv</i>
<i>Gm16197</i>	<i>Chchd10</i>
<i>Gm7550</i>	<i>Ctif</i>
<i>Kazn</i>	<i>Gm10565</i>
<i>Mir1904</i>	<i>Gm9104</i>
<i>Mir1929</i>	<i>Mir145</i>
<i>Neat1</i>	<i>Mir290</i>
<i>Prrc2b</i>	<i>Pirt</i>
<i>Rbfox2</i>	<i>Sik1</i>
<i>Spaca7</i>	<i>Tfcp2l1</i>
<i>Tmem220</i>	

To identify whether either class of ESC Super Enhancer shows a selective response to any of the shRNA targets, I looked at the effect of the shRNA depletion on the expression of Maintained and Silenced ESC Super Enhancer associated genes. Of the 123 Maintained and 77 Silenced Super Enhancer associated genes, 12 and 11 had no target probe on this microarray platform (table 6.2). For the most part, the missing genes comprise Fam genes (genes of unknown function that share highly similar DNA sequences), Gm genes (predicted genes), and genes with no known function in ESCs. Apart from these genes, *Tfcp2l1* and *Mir290* with known functions were also missing on the platform. *Tfcp2l1* acts downstream of both 2i and serum self-renewal targets.(36,299) *Mir290* plays a complex role in self-renewal

and pluripotency by regulating the Wnt pathway, polycomb proteins, and critical ESC transcription factors (see Yuan *et al.*, for a review(300)). Although unfortunate the arrays lacked target probes for some ESC Super Enhancer associated genes, as these genes form such a small proportion of the ESC Super Enhancer genes overall, one would not expect a dramatic impact on a rank-based statistic evaluating differential expression.

To narrow the list of candidate shRNA targeted molecules with a selective effect on either class of ESC Super Enhancers, I used a mean-rank gene set test²⁰ on the Maintained and Silenced Super Enhancer associated genes. This tests whether a set of genes ranks highly relative to all other genes in terms of a given statistic. In this instance, the mean-rank gene set test determines whether the t-statistics pertaining to Maintained and Silenced Super Enhancer associated genes tend toward differential expression in either direction compared to those for all other probes on the array. The outputted P-value reflects the statistical significance of differential expression. I used the Bonferroni method to adjust for multiple hypothesis testing. Figure 5.1D visualises the $-10 \times \log_{10}(\text{P-values})$ for differential expression of the genes associated with Maintained (x-axis) and Silenced (y-axis) Super Enhancers (see table 5.2 for detail). The grey lines indicate the threshold for statistical significance (P-value = 0.05; $-10 \times \log_{10}(\text{P-value}) = 13.01$). The four sets of points (grey, black, green and pink) represent shRNA target microarrays for which the differential expression of the Super Enhancer associated genes was as follows:

- Bottom left (grey): Maintained = non-significant, Silenced = non-significant;
- Top right (black): Maintained = significant, Silenced = significant;
- Bottom right (green): Maintained = significant, Silenced = non-significant;
- Top left (pink): Maintained = non-significant, Silenced = significant.

Only 12 shRNA targets showed no statistically significant change in differential expression for either Maintained or Silenced Super Enhancer associated genes. This is perhaps unsurprising

²⁰ Methods: Chapter 2.2.2.5.2

as the shRNA targets were selected “based on their relevance to gene regulation and their relatively high expression in mouse ES cells”.(297) 56 shRNA targets showed statistically significant differential expression of both Maintained and Silenced Super Enhancer associated genes. These include key naïve (*Esrrb*, *Tbx3*, *Ncoa3*), general (*Pou5f1*, *Sox2*, *Nanog*), and primed (*Eomes*, *T*) pluripotency factors.(33–36) 29 shRNA targets showed statistically significant different expression of only Maintained Super Enhancer associated genes. Notable members of this gene set include: *Pou5f1* (OCT4) and *Otx2*, which act as pioneer factors in EpiSCs;(169,170) and the epigenetic remodellers *Kdm4c* (Jmjd2c), and *Eed* and *Suz12*, two members of PRC2 complexes, which respectively demethylate H3K9 and H3K36 and methylate H3K27, and play a role in maintaining the primed state of lineage-affiliated genes in both ESCs and EpiSCs.(228) 7 shRNA targets showed statistically significant differential expression of the Silenced and not the Maintained Super Enhancer associated genes. Of these only *Klf5* and *Nanog* possess known roles in ESCs.

Finally, to identify genes that dissimilarly impact on their expression, I calculated and ranked the difference in median t-statistic (Δ median t) between Silenced and Maintained associated genes (figure 5.1E and table 5.2). I excluded the 12 arrays for which the shRNA target showed no statistically significant differential expression of either gene set. Of the 100 shRNA targets (109 included duplicates), only 6 exhibit a median $|t| > 2$. shRNAs targeting *Esrrb* (-4.83), *Nanog* (-2.91), and *Pou5f1*-B (OCT4) (-2.03) show the smallest Δ median t-statistics, meaning their inhibition reduces the expression of Silenced relative to Maintained associated genes. *Ncoa3*, the co-factor(301,302) of *Esrrb* shows the 11th smallest Δ median t-statistic (-1.25), which further supports a role for *Esrrb* as a regulator of Silenced Super Enhancers. shRNAs targeting *Otx2* (1.60) *Sox2* (2.15) and *Pou5f1*-N (OCT4) (2.69) show the largest Δ Median t-statistics, meaning their inhibition reduces the expression of Maintained relative to Silenced associated genes. To explain the conflicting response to shRNA *Pou5f1* (OCT4), I further examined how Nishiyama *et al.* inhibited its expression. After transfecting ESCs with shRNA targeted to *Pou5f1* cells were split and treated as technical replicates, named *Pou5f1*-N and

Pou5f1-B. They used *Pou5f1*-N to optimise the shRNA treatment duration, testing 48, 72, and 96 hour incubations. The two inconsistent *Pou5f1* shRNA cultures are *Pou5f1*-B (72 hr) and *Pou5f1*-N (96 hr). All 4 treatments showed a large and statistically significant downregulation of *Pou5f1* (*Pou5f1*-B: $\log_2(\text{FC})=-4.82$, adjusted P-value= 1.47×10^{-79} . *Pou5f1*-N (48hr): $\log_2(\text{FC})=-2.38$, 1.61×10^{-47} . *Pou5f1*-N (72 hr): $\log_2(\text{FC})=-3.41$, 1.68×10^{-63} . *Pou5f1*-N (96hr): $\log_2(\text{FC})=-2.85$, 3.64×10^{-55}). As the incubation time of *Pou5f1*-N increases, Super Enhancer associated genes decrease in expression (Mean: 48 hrs=-1.5, 72 hrs=-5.61, 96 hrs=-8.20; Welch Two Sample t-test P-value for 96 hrs vs 48 hrs = 0.00017). Therefore, failed transfection or inhibition of *Pou5f1* does not explain this contradiction. As such, I discounted *Pou5f1* as a candidate gene. *Nanog* displays the same discrepancy. The *Nanog*-B transfected cells show the second smallest Δ median t-statistic, however the second *Nanog* shRNA transfected cells show minimal difference (-0.20). Both microarrays show a clear and statistically significant downregulation of *Nanog* (*Nanog*-B: $\log_2(\text{FC}) = -1.13$, adjusted P-value = 2.21×10^{-36} , *Nanog*: $\log_2(\text{FC}) = -2.03$, adjusted P-value = 5.30×10^{-60}), as well as an overall decrease in Super Enhancer associated gene expression (Median expression: *Nanog*-B = -3.30, *Nanog* = -2.11). One possible explanation for the difference in expression of Super Enhancer associated genes could be that splitting the culture unevenly divided a heterogeneous pool of cells. Each cell will contain and express a slightly different amount of the shRNA vector. As such each cell's transcriptome may react differently to the loss of the targeted transcription factor. In the absence of further replicates one can't conclude how consistent the changes in gene expression may be. As the two *Nanog* targeted shRNA cell populations do not completely contradict one another, I opted to continue investigating a potential role for *Nanog* in differentially regulating Maintained and Silenced Super Enhancers. Of the remaining shRNA targets, *Esrrb* (-4.83), *Nanog* (-2.91), *Otx2* (1.60) and *Sox2* (2.15) exhibit the largest Δ median t-statistics (see figure 5.1F for a side-by-side comparison), and have well-characterised roles in pluripotent stem cell biology. Thus, they present the strongest candidates as differential regulators of Maintained (*Otx2* and *Sox2*) and Silenced (*Esrrb* and *Nanog*) Super Enhancers in ESCs.

Table 5.2 The expression of ESC Super Enhancer associated genes upon shRNA inhibition of pluripotency transcription factors.

shRNA target – the pluripotency transcription factor targeted for shRNA depletion. Median SE – the median t-statistic for differential expression of all Super Enhancer associated genes. Enrichment: the FDR corrected P-value for differential expression of Maintained and Silenced ESC Super Enhancer associated genes, the Differential Expression (DE) column indicates whether upon shRNA depletion the Maintained, Silenced, or both gene sets are statistically significantly enriched for differential expression. Median t for DE: the median t-statistics for differential expression of Maintained and Silenced ESC Super Enhancer associated genes, the Δ Median t = median Silenced t – median Maintained t. The table is sorted by the Δ Median t column.

shRNA target	Median SE	Enrichment		DE	Median t for DE		
		Maintained	Silenced		Maintained	Silenced	Δ Median t
Esrrb	-10.23	2.66E-04	4.36E-04	Both	-8.92	-13.75	-4.83
Nanog-B	-3.30	2.66E-04	4.36E-04	Both	-2.37	-5.27	-2.91
Pou5f1-B	-8.29	2.66E-04	7.03E-04	Both	-7.63	-9.66	-2.03
Btbd14b	-1.36	9.13E-03	4.36E-04	Both	-0.74	-2.54	-1.80
Mycl1	-1.33	6.41E-04	4.36E-04	Both	-1.18	-2.67	-1.48
Etv1	-0.54	8.63E-01	3.63E-03	Silenced	-0.06	-1.51	-1.44
Ctbp2	-1.58	2.66E-04	4.36E-04	Both	-1.30	-2.71	-1.41
Tbx3	-2.35	2.66E-04	4.36E-04	Both	-1.46	-2.79	-1.33
Ncoa3	-2.17	2.66E-04	4.36E-04	Both	-1.57	-2.82	-1.25
Aff1	-2.80	2.66E-04	7.03E-04	Both	-2.49	-3.42	-0.93
Fbxo15	-1.11	3.04E-01	5.78E-03	Silenced	-1.01	-1.92	-0.90
Foxg1	-0.70	6.50E-01	1.21E-03	Silenced	-0.25	-1.14	-0.90
Tcfap4	-2.01	6.41E-04	7.03E-04	Both	-1.57	-2.43	-0.86
Tle4	-2.06	2.66E-04	7.03E-04	Both	-2.00	-2.66	-0.67
Mybl2	-1.25	1.23E-03	6.38E-03	Both	-1.07	-1.72	-0.65
D1pas1	-0.72	9.13E-03	4.36E-04	Both	-0.50	-1.09	-0.59
Mtf2	-3.30	2.66E-04	4.36E-04	Both	-3.20	-3.77	-0.57
Cdyl	-0.84	3.54E-02	9.91E-04	Both	-0.70	-1.24	-0.54
Foxd3	-0.67	9.13E-03	3.13E-02	Both	-0.61	-1.09	-0.48
Gbx2	-1.51	1.07E-03	4.36E-04	Both	-1.43	-1.89	-0.46
Klf4	-2.06	2.66E-04	4.36E-04	Both	-1.88	-2.28	-0.40
Dnmt3b	-1.94	4.45E-04	1.21E-03	Both	-1.75	-2.14	-0.39
Foxp1	-1.13	4.45E-04	2.03E-03	Both	-0.99	-1.37	-0.38
Sall4	-4.14	2.66E-04	4.36E-04	Both	-4.05	-4.36	-0.32
Ctnnb1	-1.81	4.45E-04	4.36E-04	Both	-1.78	-2.07	-0.29
Trim24	-1.43	2.66E-04	2.03E-03	Both	-1.23	-1.51	-0.28
Eomes	-0.52	1.72E-02	1.66E-02	Both	-0.45	-0.71	-0.26
Nanog	-2.11	5.64E-02	4.36E-04	Silenced	-2.02	-2.21	-0.20
Elf2	-0.30	9.70E-02	1.17E-02	Silenced	-0.27	-0.45	-0.18
Sall1	-1.17	2.66E-04	1.86E-03	Both	-1.14	-1.30	-0.17
Sub1	-1.51	2.66E-04	4.36E-04	Both	-1.48	-1.63	-0.15
Tlk1	-1.49	2.66E-04	1.86E-03	Both	-1.46	-1.59	-0.13
Hmga1	-1.22	7.79E-04	1.26E-02	Both	-1.13	-1.24	-0.11
Tbl1x	-0.67	2.26E-02	1.05E-01	Maintained	-0.67	-0.78	-0.11
Dppa2	-1.57	2.66E-04	2.48E-03	Both	-1.51	-1.59	-0.08
Mkrrn1	-1.12	8.89E-03	5.86E-02	Maintained	-1.10	-1.16	-0.06
Pml	-1.25	4.45E-04	8.13E-03	Both	-1.23	-1.28	-0.05
Cdyl2	-0.20	1.04E-02	3.49E-01	Maintained	-0.20	-0.21	0.00
Fgfbp1	-1.20	2.66E-04	2.06E-02	Both	-1.20	-1.20	0.00
Inpp1	-0.94	7.79E-04	5.87E-03	Both	-0.94	-0.94	0.00
Tcl1	-1.42	2.66E-04	1.72E-03	Both	-1.43	-1.39	0.03
Zfp57	-1.05	1.58E-03	6.92E-02	Maintained	-1.05	-1.00	0.04
Trim28	-1.16	1.34E-03	4.03E-03	Both	-1.16	-1.12	0.05
Pou5f1-N (72 hr)	-3.63	2.66E-04	4.36E-04	Both	-3.66	-3.59	0.07
Klf5	-1.04	2.09E-01	5.78E-03	Silenced	-1.07	-0.96	0.11
Whsc2	0.55	1.86E-01	9.34E-03	Silenced	0.50	0.62	0.13
Klf2	-1.86	2.66E-04	4.36E-04	Both	-1.91	-1.77	0.14

<i>Gli1</i>	-1.06	1.88E-02	1.44E-01	Maintained	-1.06	-0.91	0.15
<i>Eed</i>	-0.71	2.46E-03	1.15E-01	Maintained	-0.78	-0.62	0.15
<i>Ebf1</i>	-0.55	1.07E-03	9.91E-04	Both	-0.73	-0.51	0.21
<i>Lin28</i>	0.47	2.46E-02	1.32E-01	Maintained	0.30	0.52	0.22
<i>Trp53</i>	-1.16	7.79E-04	7.03E-04	Both	-1.17	-0.85	0.33
<i>Zfp281</i>	-0.34	2.66E-04	4.21E-01	Maintained	-0.45	-0.12	0.33
<i>Pdlim1</i>	-0.47	3.26E-02	9.61E-01	Maintained	-0.53	-0.18	0.35
<i>Rpap1</i>	-1.44	1.58E-03	1.02E-01	Maintained	-1.69	-1.31	0.37
<i>Jmjd1a</i>	-1.25	1.46E-03	1.72E-03	Both	-1.40	-0.98	0.42
<i>Zscan4c</i>	-0.72	1.34E-03	5.86E-02	Maintained	-0.92	-0.48	0.43
<i>Aire</i>	-1.14	4.45E-04	4.19E-01	Maintained	-1.35	-0.91	0.44
<i>Ash2l</i>	-0.85	4.39E-03	8.18E-02	Maintained	-0.94	-0.49	0.46
<i>Smad4</i>	-2.05	2.66E-04	1.83E-02	Both	-2.30	-1.83	0.47
<i>Dppa5a</i>	-2.09	2.66E-04	4.36E-04	Both	-2.23	-1.68	0.56
<i>Cdc5l</i>	-2.97	2.66E-04	4.36E-04	Both	-3.11	-2.46	0.65
<i>Id1</i>	-1.00	1.01E-02	4.86E-01	Maintained	-1.16	-0.45	0.70
<i>Sox15</i>	-0.85	9.13E-03	5.93E-01	Maintained	-1.23	-0.51	0.71
<i>Dppa3</i>	-0.87	9.56E-04	5.87E-03	Both	-1.01	-0.29	0.72
<i>Nr6a1</i>	-0.77	1.07E-03	6.29E-01	Maintained	-1.04	-0.31	0.73
<i>Elf2</i>	-2.27	2.66E-04	4.36E-04	Both	-2.43	-1.68	0.74
<i>Pou5f1-N (48 hr)</i>	-1.48	2.66E-04	4.36E-04	Both	-1.59	-0.85	0.75
<i>Zic3</i>	-2.18	1.07E-03	5.87E-03	Both	-2.39	-1.64	0.75
<i>Sap30</i>	-1.56	2.66E-04	1.21E-03	Both	-2.03	-1.25	0.77
<i>Dnmt3l</i>	-1.68	2.66E-04	1.26E-02	Both	-1.73	-0.96	0.78
<i>Prdm5</i>	-0.80	4.45E-04	1.23E-01	Maintained	-1.09	-0.26	0.83
<i>Hells</i>	-0.69	1.34E-03	5.95E-01	Maintained	-1.24	-0.33	0.90
<i>Gtf2i</i>	-1.13	2.66E-04	7.47E-03	Both	-1.60	-0.69	0.91
<i>Zfp42</i>	-0.96	2.46E-03	1.58E-01	Maintained	-1.13	-0.21	0.92
<i>Parp1</i>	-0.77	4.45E-04	1.58E-01	Maintained	-1.11	-0.18	0.92
<i>T</i>	-1.67	2.66E-04	7.03E-04	Both	-1.98	-0.97	1.02
<i>Utf1</i>	-1.55	2.66E-04	3.81E-02	Both	-1.93	-0.90	1.02
<i>Phc1</i>	-1.90	2.66E-04	5.93E-01	Maintained	-2.07	-1.03	1.04
<i>Kdm4c</i>	-1.20	2.66E-04	6.18E-02	Maintained	-1.69	-0.62	1.08
<i>Suz12</i>	-0.39	3.54E-02	9.55E-01	Maintained	-0.97	0.15	1.12
<i>Nr0b1B</i>	-1.24	7.79E-04	4.95E-02	Both	-1.45	-0.31	1.14
<i>Tcfcp2l1</i>	-1.45	2.66E-04	4.17E-03	Both	-1.83	-0.63	1.19
<i>Mycn</i>	-0.63	5.08E-03	9.88E-01	Maintained	-1.50	-0.26	1.24
<i>Nr5a2</i>	-0.43	5.81E-03	1.06E-01	Maintained	-0.75	0.51	1.26
<i>Mbd3</i>	-2.09	2.66E-04	4.36E-04	Both	-2.45	-1.03	1.42
<i>Tcfap2c</i>	-1.24	2.66E-04	1.42E-02	Both	-1.79	-0.36	1.43
<i>Rest</i>	-0.97	7.79E-04	8.69E-01	Maintained	-1.41	0.09	1.51
<i>Zfp206</i>	-0.35	2.66E-04	2.38E-01	Maintained	-0.95	0.60	1.55
<i>Otx2</i>	-1.67	2.66E-04	1.28E-01	Maintained	-2.40	-0.80	1.60
<i>Sox2</i>	-5.91	2.66E-04	4.36E-04	Both	-5.92	-3.77	2.15
<i>Pou5f1-N (96 hr)</i>	-4.02	1.46E-03	5.95E-01	Maintained	-5.42	-2.73	2.69

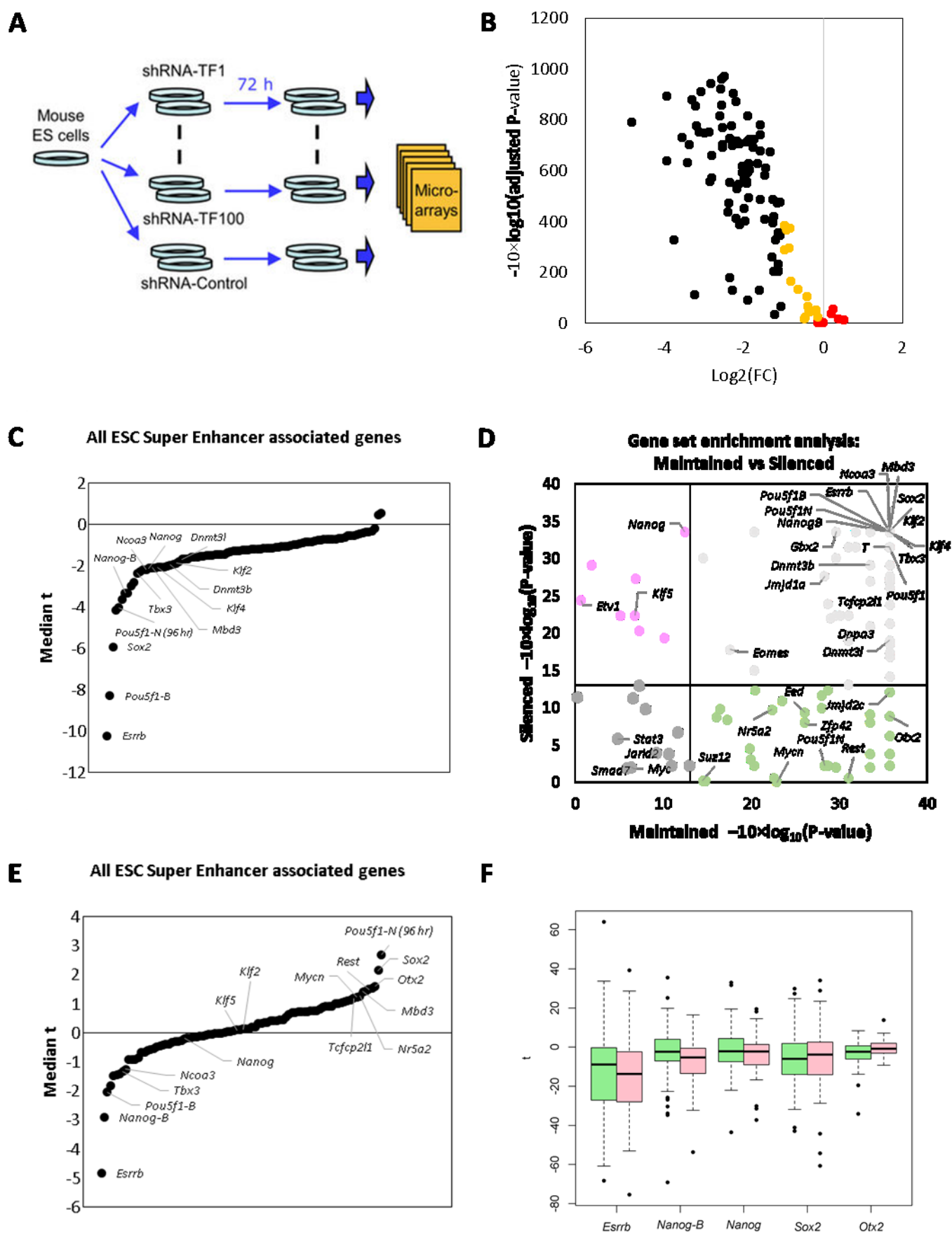


Figure 5.1 Identifying candidate transcription factors that selectively regulate PU and DM ESC Super Enhancer activity using a publicly available shRNA screen.

A. A schematic of Nishiyama et al.'s experimental design adapted from their publication.[3] Mouse ESCs were transfected with shRNA targeting 100 key ESC transcription factors. After 72 hours of puromycin selection, cells were harvested, their RNAs extracted, and applied to microarrays. ESCs transfected with an empty vector were used as control. **B.** The $\log_2(\text{fold-change})$ in expression of the genes targeted with shRNAs plotted against the $-10 \times \log_{10}(\text{adjusted } P\text{-value})$, representing the statistical significance of their differential expression. Red = genes with a $\log_2(\text{fold-change}) > 0$ or an adjusted $P\text{-value} > 0.05$. These microarrays were excluded from further analysis. Yellow = genes with a $-1 < \log_2(\text{fold-change}) < 0$ and an adjusted $P\text{-value} < 0.05$. Black = genes with a $\log_2(\text{fold-change}) < -1$ and an adjusted $P\text{-value} < 0.05$. **C.** The median t -statistic for differential expression of ESC Super Enhancer associated genes upon shRNA knock down of the panel of 100 key ESC transcription factors. The labelled points highlight the naive (*Esrrb*, *Klf2*, *Klf5*) and general (*Pou5f1*, *Sox2*, *Nanog*) pluripotency transcription factors, and genes related to DNA methylation (*Mbd3*, *Dnmt3b*, *Dnmt3l*). **D.** Competitive mean-rank gene set test P -values indicating whether Maintained and Silenced ESC Super Enhancer associated genes are statistically significantly differentially expressed upon shRNA knock down of the panel of 100 key ESC transcription factors. The P -values are adjusted for multiple hypothesis testing using the Bonferroni method and given as $-10 \times \log_{10}(\text{adjusted } P\text{-value})$. The grey lines indicate a significance threshold of $-10 \times \log_{10}(\text{adjusted } P\text{-value}) > 13.01$ (i.e. $P < 0.05$). The colour key indicates those shRNA experiments where Silenced (pink), Maintained (light green), neither (grey) or both (black) classes ESC Super Enhancer are statistically significantly differentially expressed. The highlighted genes relate to naive (*Esrrb*, *Ncoa3*, *Gbx2*, *Klf2*, *Klf4*, *Klf5*, *Tbx3*, *Dppa3*, *Zfp42*), general (*Pou5f1*, *Sox2*, *Nanog*), and primed (*Otx2*, *Eomes*) pluripotency factors, histone modifications (*Jarid2*, *Jmjd2*, *Eed*, *Suz12*), DNA methylation (*Dnmt3b*, *Dnmt3l*, *Mbd3*), or genes that differentially expressed in only one class of Super Enhancer associated genes. **E.** The difference in median t -statistic for the differential expression of Silenced versus Maintained ESC Super Enhancer associated genes of 100 key ESC transcription factors. The labelled points highlight the naive (*Esrrb*, *Klf2*, *Klf5*, *Tbx3*) and general (*Pou5f1*, *Sox2*, *Nanog*) pluripotency transcription factors, and genes related to DNA methylation (*Mbd3*), or genes that differentially expressed in only one class of Super Enhancer associated genes. **F.** Boxplots showing the t -statistics for the differential expression of Silenced (pink) and Maintained (light green) ESC Super Enhancer associated genes upon shRNA inhibition of key ESC transcription factors. These plots shown highlight the shRNA targets that induce the largest differences in median t -statistic between the two classes of Super Enhancer.

5.2.2 ERRE and KLF motifs appear enriched at DM Super Enhancelets

The analysis of Nishyama *et al.*'s ESC shRNA screen described in chapter 5.2.1 raised ESRRB, NANOG, SOX2 and OTX2 as potential regulators of Maintained and Silenced ESC Super Enhancers. Each of these molecules function as transcription factors, which bind to a specific DNA sequence within a regulatory DNA region to exert an effect on transcription. As such, if they do regulate Maintained or Silenced Super Enhancers their DNA binding motif may appear enriched within PU or DM Super Enhancelets. To examine the DNA binding motifs present at ESC Super Enhancers I employed HOMER(212) to find enriched motifs. In the publication defining ESC Super Enhancers, Whyte *et al.* reported that constituent enhancers relative to randomised genomic background showed statistically significant enrichment of the DNA binding motifs for OCT4 ($P = 9.19 \times 10^{-64}$), SOX2 ($P = 3.01 \times 10^{-67}$), NANOG ($P = 9.46 \times 10^{-17}$), KLF4 ($P = 4.33 \times 10^{-6}$) and ESRRB ($P = 2.55 \times 10^{-84}$), and that the CTCF motif showed no significant enrichment ($P = 0.45$).⁽¹⁷⁹⁾ To confirm that HOMER could replicate these results, I tested for enrichment at the unmethylated constituent enhancers defined in ESCs compared to randomised genomic background and interstitial Persistently Methylated (PM) subregions. I included the comparison to PM as they provide a more biologically relevant control regions than artificially generated randomised background, as they form part of the same genomic regions (i.e. ESC Super Enhancers). Figure 5.2A and table 5.2 detail the $-10 \times \log_{10}(Q\text{-values})$, where the Q-value is the FDR corrected P-value, for each DNA binding motif (the dotted grey line indicates a significance threshold of $Q < 0.05$). As $\log(0) = -\infty$, to complete the log transformation, I replaced Q values equal to 0 with the smallest non-zero value (0.0001). HOMER successfully identified the OCT4-SOX2-NANOG, SOX2, NANOG, KLF4 and ESRRB binding motifs as significant, and the CTCF motif as non-significant as previously reported.⁽¹⁷⁹⁾ As well as the ESRRB motif, I included the highly similar ERRA and NR5A2 motifs. Each of these appeared enriched at unmethylated constituent enhancers. In chapter 5.2.1, we saw that shRNA depletion of numerous pluripotency-related transcription factors selectively led to the differential expression of Maintained or Silenced ESC Super Enhancer

associated genes. While shRNA inhibition of *Esrrb* and *Nanog* (Silenced), and *Sox2* and *Otx2* (Maintained), showed the largest

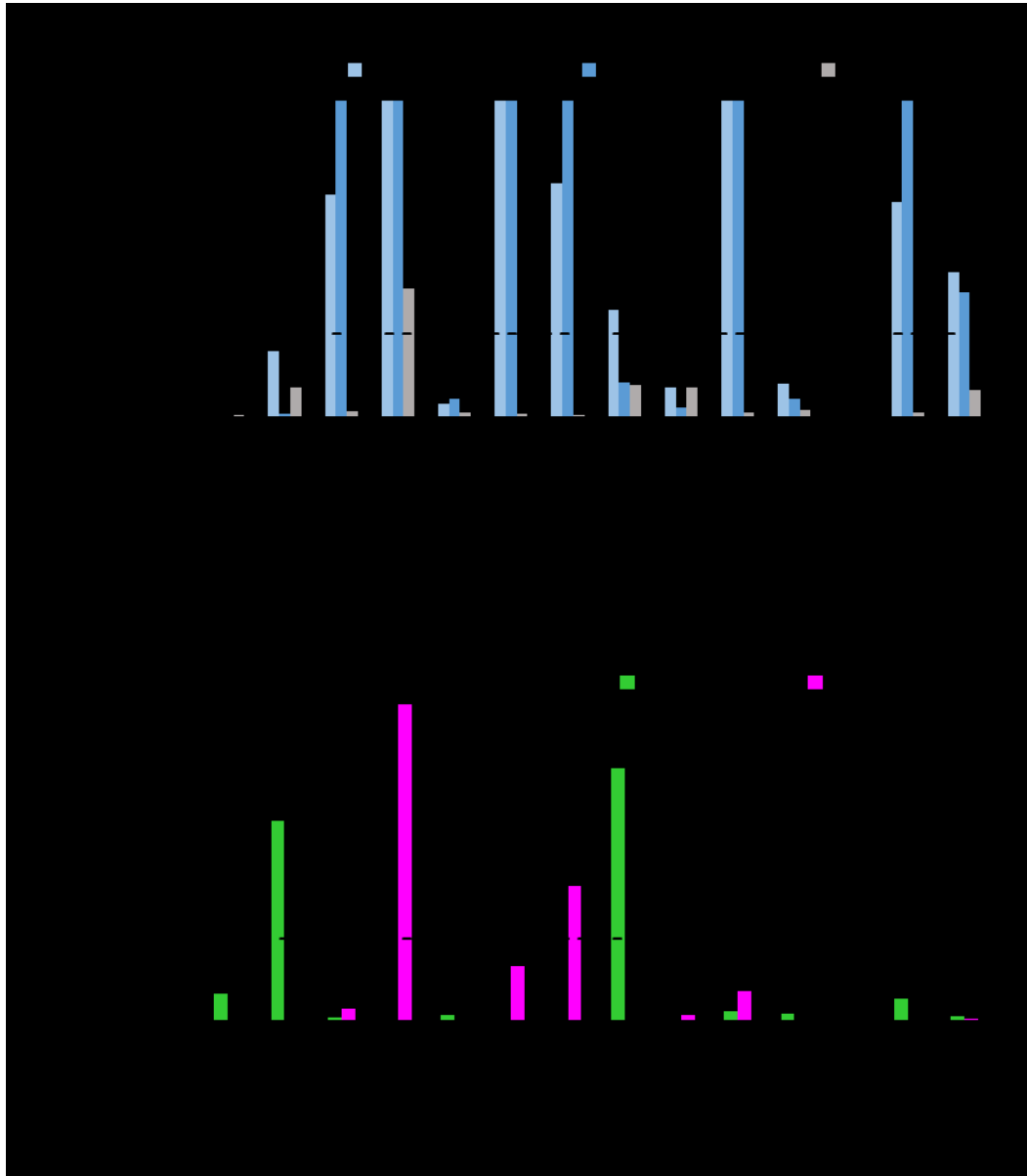


Figure 5.2 DNA binding motifs selectively enriched at PU and DM Super Enhancers.

I attempted to identify candidate transcription factors that selectively regulate PU and DM ESC Super Enhancers by the relative enrichment of their DNA binding motifs. HOMER motif analysis of the candidate PU and DM regulators identified by shRNA screen. The barplots show the $-10 \times \log_{10}(\text{adjusted } P\text{-value})$ for enrichment in the regions of interest relative to the given background. The black dotted line indicates a significance threshold of $-10 \times \log_{10}(\text{adjusted } P\text{-value}) > 13.0103$ (i.e. $P < 0.05$).

Table 5.3 HOMER motif analysis of ESC Super Enhancelets.

The first two columns detail the identity of the protein and motif. The remaining five columns list the $-10 \times \log_{10}(\text{Q-value})$ (where the Q-value is the FDR corrected P-value) for enrichment of that motif in ESC Super Enhancer subregions. Protein – the protein that binds to the identified motif, HOMER motif name – the name of the motif as indexed by HOMER. Unmeth – all Super Enhancelets, BG – randomised genomic background, PM – Persistently Methylated interstitial subregions, PU – Persistently Unmethylated Super Enhancelets, DM – Differentially Methylated Super Enhancelets. The figures highlighted in green indicate a significance threshold of $-10 \times \log_{10}(\text{adjusted P-value}) > 13.0103$ (i.e. $P < 0.05$).

Protein	HOMER motif name	Unmeth vs BG	Unmeth vs PM	PM vs BG	PU vs DM	DM vs PU
BRACHYURY	Brachyury(T-box)/Mesoendoderm-Brachyury-ChIP-exo(GSE54963)/Homer	0.00	0.00	0.28	4.21	0.00
CTCF	CTCF(Zf)/CD4+CTCF-ChIP-Seq(Barski et al.)/Homer	10.29	0.45	4.57	31.55	0.00
ERRA	Erra(NR)/HepG2-Erra-ChIP-Seq(GSE31477)/Homer	35.23	50.00	0.81	0.38	1.93
ESRRB	Esrrb(NR)/mES-Esrrb-ChIP-Seq(GSE11431)/Homer	50.00	50.00	20.22	0.00	50.00
ETV1	ETV1(ETS)/GIST48-ETV1-ChIP-Seq(GSE22441)/Homer	2.02	2.71	0.59	0.93	0.00
KLF4	Klf4(Zf)/mES-Klf4-ChIP-Seq(GSE11431)/Homer	50.00	50.00	0.41	0.00	8.60
KLF5	KLF5(Zf)/LoVo-KLF5-ChIP-Seq(GSE49402)/Homer	36.99	50.00	0.13	0.00	21.25
NANOG	Unknown-ESC-element(?)/mES-Nanog-ChIP-Seq(GSE11724)/Homer	16.84	5.40	4.93	40.00	0.00
N-MYC	n-Myc(bHLH)/mES-nMyc-ChIP-Seq(GSE11431)/Homer	4.55	1.41	4.57	0.00	0.83
OCT4	OCT4-SOX2-TCF-NANOG(POU,Homeobox,HMG)/mES-Oct4-ChIP-Seq(GSE11431)/Homer	50.00	50.00	0.59	6.60	0.00
OTX2	Otx2(Homeobox)/EpiLC-Otx2-ChIP-Seq(GSE56098)/Homer	5.24	2.71	1.00	1.05	0.00
REST	REST-NRSF(Zf)/Jurkat-NRSF-ChIP-Seq/Homer	0.00	0.00	0.00	0.00	0.00
SOX2	Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer	33.98	50.00	0.52	3.49	0.00

median shift in differential expression of the genes associated with the two classes of ESC Super Enhancer, other transcription factors showed smaller, but still selective effects on these classes of genes based on the gene set enrichment analysis, some of which have indexed DNA binding motifs, namely BRACHYURY (*T*), ETV1, KLF5, N-MYC, OTX2, REST and TCFCP2L1. Of these, only KLF5 and TCFCP2L1 appear statistically significantly enriched at the unmethylated constituent enhancers. This doesn't disprove that the remaining transcription factors do not selectively regulate Maintained or Silenced ESC Super Enhancer associated genes, but if they do they it may be through protein-protein interactions.

To test for motif enrichment within PU or DM Super Enhancelets, I compared PU against DM, and DM against PU (figure 5.2B, table 5.3). If OTX2 and SOX2 regulates PU and ESRRB and NANOG regulate DM Super Enhancelets, one might observe an enrichment of their motifs at these regions. ESRRB indeed appears enriched at DM Super Enhancelets. KLF5 also appears enriched at DM compared to PU Super Enhancelets. The treatment of ESCs with shRNA targeting KLF5 causes the statistically significant differential expression of Silenced associated genes only. As KLF2, KLF4, and KLF5 compete for the same DNA binding sites,(303) it's not clear which of the KLF family this points to. Regardless, this further evidences a possible role for ESRRB and the KLF family in regulating the activity of DM Super Enhancelets. With regards to the PU Super Enhancelets, neither OTX2 nor SOX2 show an enrichment in their DNA binding motifs, however NANOG and CTCF do. The enrichment of the NANOG DNA binding motif at PU relative to DM Super Enhancelets conflicts with the identification of NANOG as the second largest driver of differential expression of Silenced compared to Maintained Super Enhancer associated genes. Recently, Ren *et al.* reported that CTCF binding correlates with enhancer activity and enhancer-promoter interactions. They concluded CTCF stabilised these interactions.(304) If CTCF does bind more densely to PU compared to DM Super Enhancelets, it may contribute to the continued expression in EpiSCs of Maintained Super Enhancer associated genes by stabilising the interactions of PU Super Enhancelets to their promoters.

In summary, the enrichment of the DNA binding motifs for ESRRB and KLF5 at DM and CTCF at PU evidences a potential regulatory role at these regions. The enrichment of the CTCF binding motif at PU was surprising given that Whyte *et al.* reported no enrichment of the CTCF binding motif at ESC Super Enhancers.⁽¹⁷⁹⁾ This potentially suggests a newly uncovered role for CTCF at ESC Super Enhancers, specifically PU Super Enhancelets. OTX2 and SOX2 may yet play a selective role at PU via indirect (i.e. protein-protein, rather than protein-DNA) recruitment by other factors.

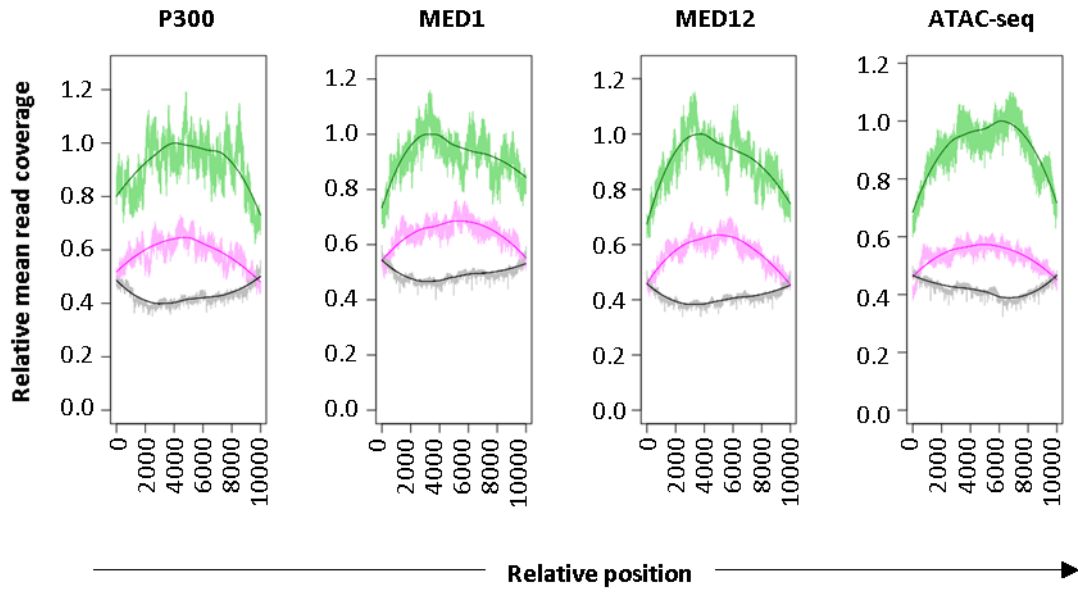
5.2.3 ESRRB and NCOA3 selectively bind to DM ESC Super Enhancelets at high density relative to other naïve and primed pluripotency factors

In chapter 4.2.6 I reported that in ESCs key pluripotency transcription factors bind at higher density to PU compared to DM ESC Super Enhancelets (figure 4.6). If a transcription factor preferentially regulates the activity of one type of Super Enhancelet over the other one might expect to see a different pattern to its binding density at these regions. To see if any of these candidate factors show an altered binding density profile, I calculated their relative mean ChIP-seq read coverage at PU, DM and PM ESC Super Enhancer subregion.²¹ I included the transcriptional co-factors H3K27 acetylase P300 and the Mediator complex members MED1 and MED12 along with ATAC-seq as references for the general pattern of transcription factor binding and chromatin accessibility at these regions. As figure 5.3 shows P300, MED1 and MED12 all bind at higher density to PU relative to DM ESC Super Enhancelets. The ATAC-seq experiment recapitulates the P300, MED1 and MED12 pattern, indicating that PU compared to DM chromatin is more accessible. However, it is unclear whether the increased transcription factor binding is cause or consequence of the increased chromatin accessibility.

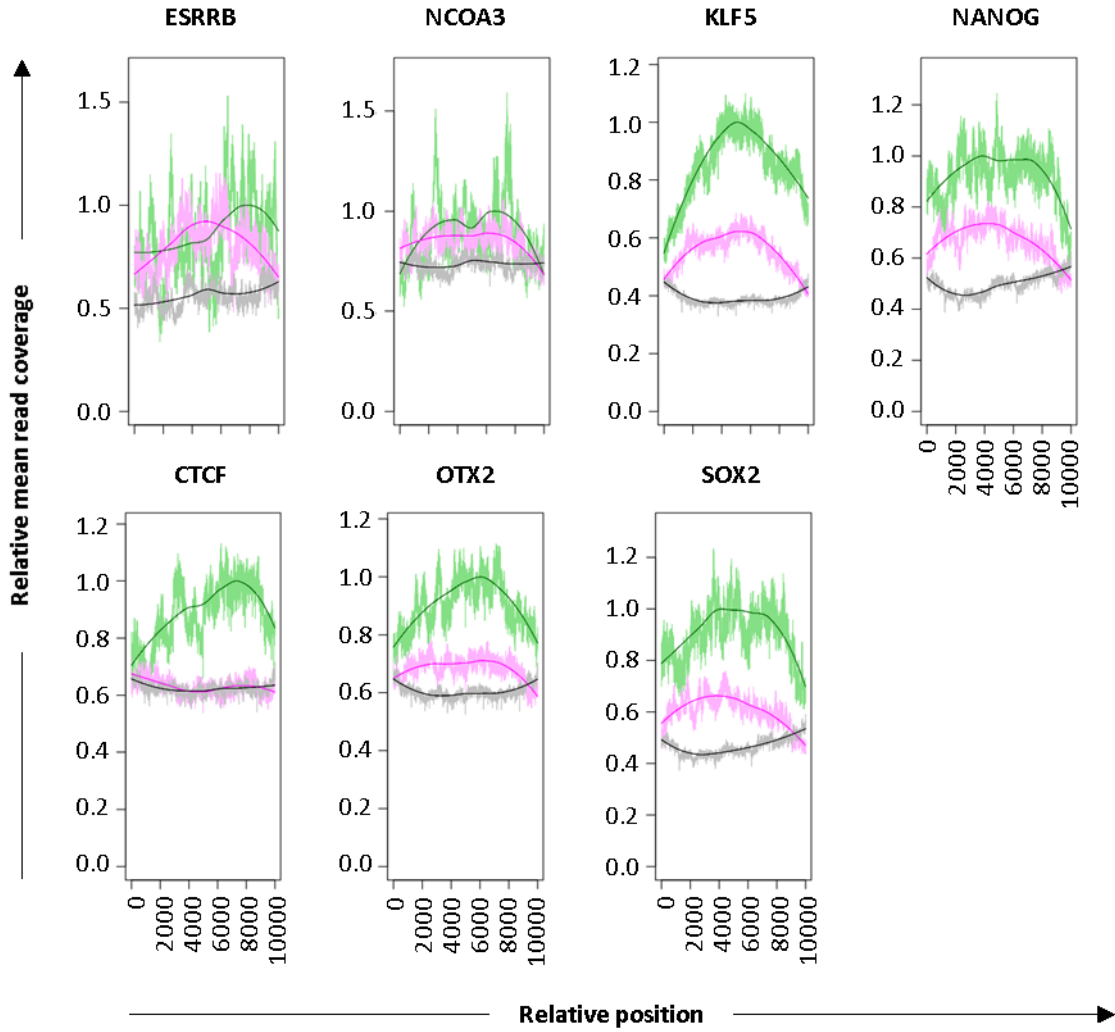
With regards to the candidate differential regulators, only ESRRB and NCOA3 diverge from the binding density profiles of the reference transcriptional co-factors. Despite the increased

²¹ Methods: Chapter 2.2.2.4.2.2

A



B



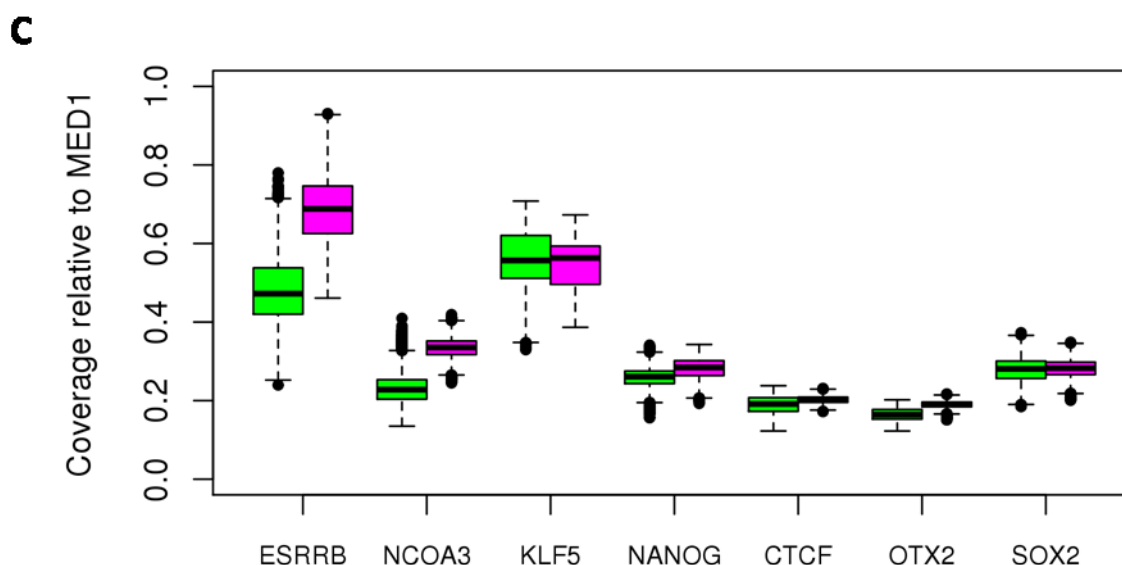


Figure 5.3 Protein binding density at PU and DM Super Enhancelets of identified candidate regulators using publicly available ChIP-seq data.

Metagenes showing the mean ChIP-seq read coverage at PU, DM and PM ESC Super Enhancer subregions for **A.** P300, Med1, Med12 and ATAC-seq, and **B.** Esrrb, Ncoa3, Klf5, Nanog, CTCF, Otx2 and Sox2. ChIP-seq bam files were normalised for sequencing depth and background. The mean read coverage values at each region set was calculated using the `plotCoverage` function (methods: chapter 2.2.2.4.2.2). The resulting mean read coverage scores were scaled to the maximal value. The darker lines show the smoothed fit of the data generated by loess regression. **C.** Boxplots summarising the read coverage of the candidate regulators at PU and DM Super Enhancelets scaled to that of MED1. DM, Differentially Methylated (magenta); PU, Persistently Methylated (green) and PM, Persistently Methylated (grey).

chromatin accessibility of PU compared to DM Super Enhancelets, they appear similarly enriched at these regions whilst KLF5, NANOG, OTX2, SOX2 and CTCF recapitulate the PU>DM>PM binding profile of P300, MED1 and MED12. To adjust for this bias in transcription factor binding, I recalculated the sequencing read density relative to MED1, the transcription factor upon which Whyte *et al.* based the definition of Super Enhancers (figure 5.3C). When corrected for chromatin accessibility, ESRRB and NCOA3 show a relative enrichment at DM compared to PU Super Enhancelets evidencing a specific regulatory function of these factors at DM regions. Of the remaining candidate transcription factors, both those associated with naïve (KLF5 and Nanog) and general/primed (SOX2 and OTX2) pluripotency show similar binding densities at PU and DM. This also holds true for CTCF.

5.2.4 Meta-analysis of microarray data evidences that Silenced ESC Super Enhancer associated genes are sensitive to the loss of ESRRB

The evidence collected so far suggests that ESRRB selectively regulates DM Super Enhancelets (figure 5.4). If this is the case, depletion of ESRRB would reduce transcription of genes associated with Silenced ESC Super Enhancers. To test this, I decided to compare the transcription candidate Maintained and Silenced ESC Super Enhancer associated genes in ESCs bearing a homozygous deletion of *Esrrb* and their wild-type counterparts (henceforth ESC^{*Esrrb*-/-} and ESC^{*Esrrb*+/+}). (305) NCOA3 acts as a co-activator of ESRRB in ESCs. To confirm that any observed differential expression of ESC Super Enhancer associated genes results from the lack of ESRRB activity, we also assayed ESC^{*Ncoa3*-/-} and ESC^{*Ncoa3*+/+} (unpublished, provided by Prof Austin Cooney). I initially identified NANOG as a candidate selective regulator of DM, however in the ensuing DNA binding motif and protein binding density analyses it failed to show a consistent preference for either set of Super Enhancelets. I conclude that NANOG does not selectively regulate PU or DM ESC Super Enhancelets. Therefore, to ensure that any selective differential expression of Silenced or Maintained ESC Super Enhancer

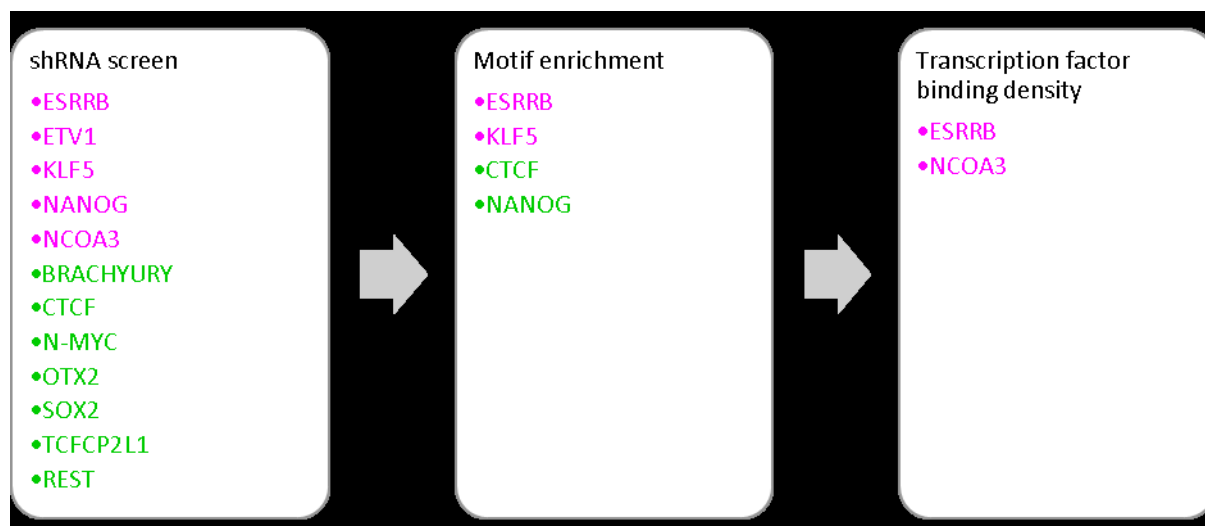
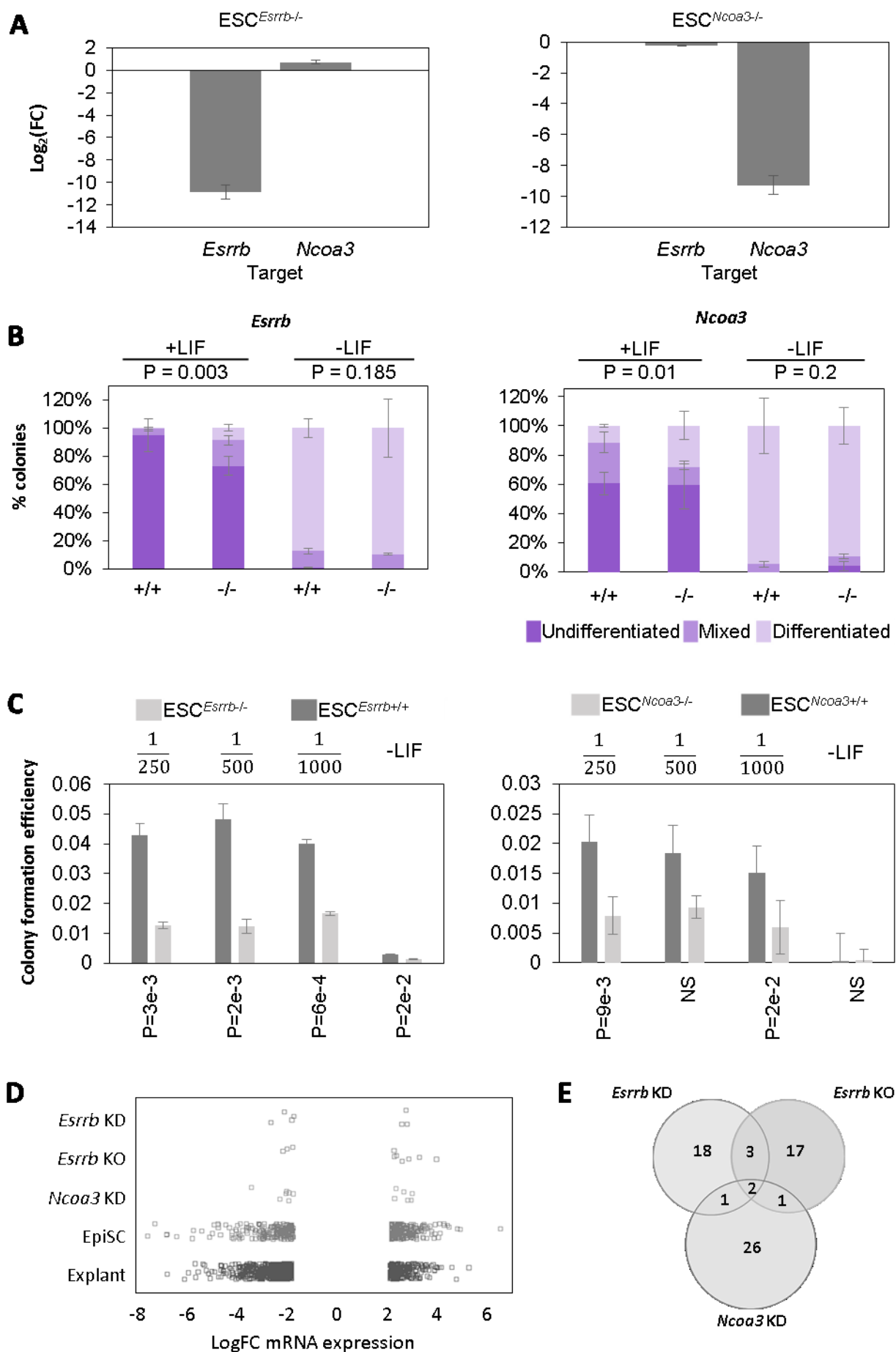


Figure 5.4 Candidate differential regulators of ESC Super Enhancelets.

Schematic summarising the candidate differential regulators of PU (green) and DM (pink) Super Enhancelets identified by shRNA screen and whether their DNA binding motifs and binding density were found to selectively enrich at PU or DM.



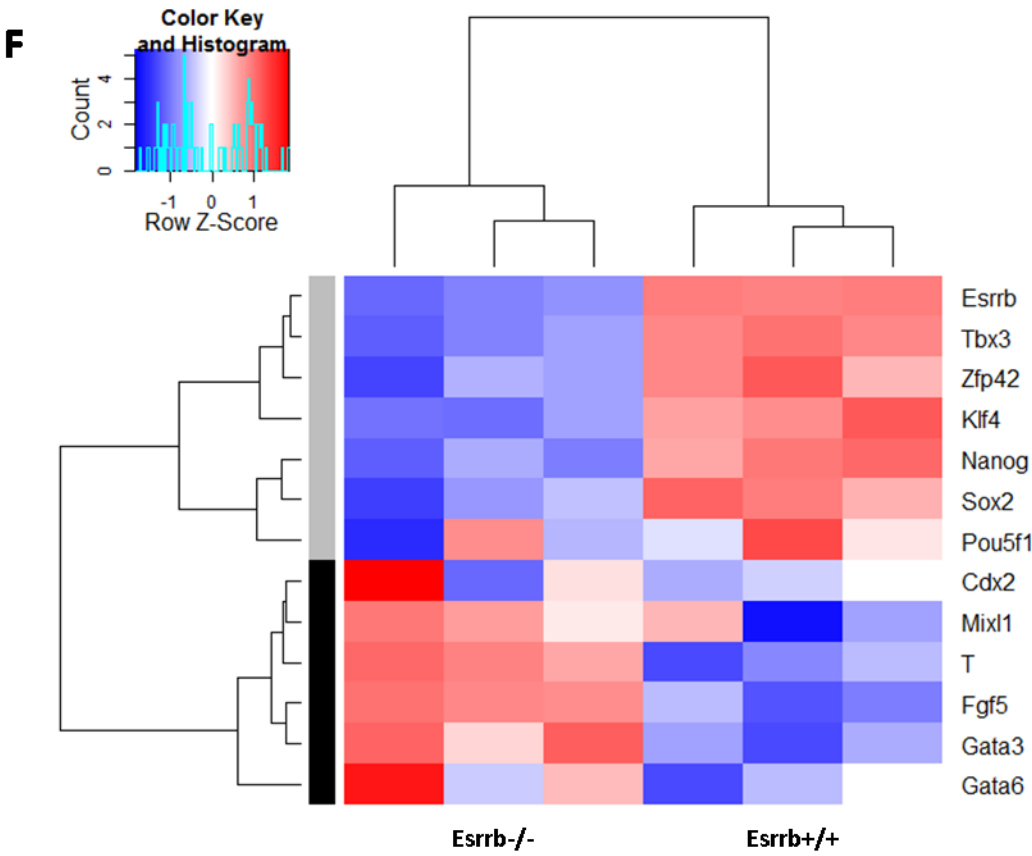
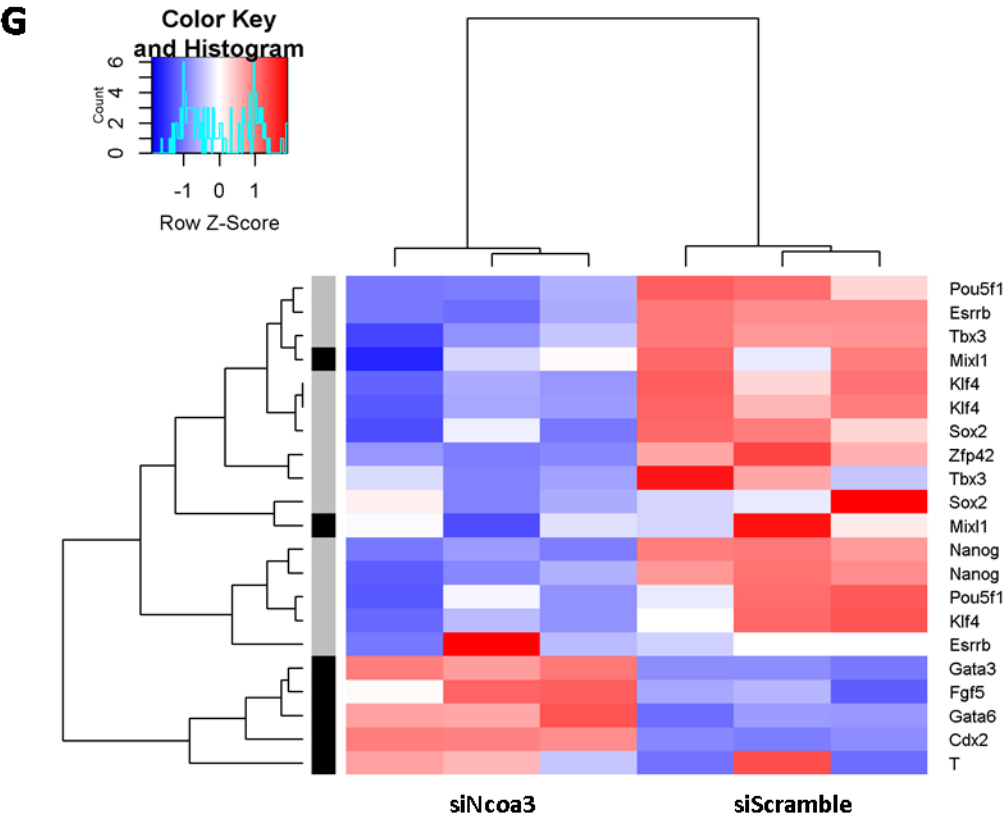


Figure 5.5 Phenotypic and functional characterisation of ESCs bearing homozygous deletions of Esrrb and Ncoa3.

A. Barplots of RT-qPCR data showing the expression of *Esrrb* and *Ncoa3* in $ESC^{Esrrb-/-}$ and $ESC^{Ncoa3-/-}$ lines compared to their wild-type controls. Data were normalised to the housekeeping genes *S17* and *L19* and expressed as $\log_2(2^{-\Delta\Delta CT})$. Three passages were collected per cell line and each qPCR was performed in triplicate. The error bars represent standard error from the mean (SEM). **B.** Stacked barplots showing the differentiation states of colonies arising from the null and wild-type control cell lines grown in the presence (1:500) and absence of LIF. Cells were seeded at low density (two wells per cell line), cultured for 8 days, after which they were stained with alkaline phosphatase and the colonies counted. The differentiation state of each colony was assigned as follows: <20% stained = differentiated; 20-90% stained = mixed; >90% stained = undifferentiated. Each pair of cell lines underwent three self-renewal assays from successive passages. The *P*-values result from unpaired *t*-tests of the percentage of undifferentiated colonies arising from the two cell lines. The error bars represent SEM. **C.** Barplots showing the colony formation efficiency of the null and wild-type control lines grown in the presence of varying concentrations of LIF. Cells were grown in the same manner as for a self-renewal assay, with the addition of two wells of each cell line per concentration of LIF. Colony formation efficiency was calculated as the number of undifferentiated and mixed colonies divided by the number of cells seeded. Each cell line underwent three colony formation efficiency assays. The *P*-values result from unpaired *t*-tests of the number of colonies arising from the two cell lines. The error bars represent SEM. **D.** Categorical scatterplot showing the $\log_2(\text{fold-change})$ in differential expression from five microarray experiments: ESCs treated with shRNA targeting *Esrrb* compared to ESCs treated with shRNA targeting GFP(307); $ESC^{Esrrb-/-}$ compared to their wild-type control(305); ESCs treated with shRNA targeting *Ncoa3* compared to ESCs treated with a scrambled shRNA(301); epiblast derived from the E5.5 embryo and cultured as EpiSCs compared to ESCs(5); epiblast derived from the E5.5 embryo and applied directly to microarray compared to ESCs(5). Only probes with $|\log_2(\text{fold-change})| > 2$ are plotted. **E.** Venn diagram showing the number of common genes with a $|\log_2(\text{fold-change})| > 2$ in the *Esrrb* and *Ncoa3* depleted cell lines. **F-G.** Heatmaps showing the $\log_2$ transformed and quantile normalised probe intensity values of naïve and general (grey) and primed (black) pluripotency associated genes in $ESC^{Esrrb-/-}$ compared to their wild-type control (**F**) and $ESC^{Ncoa3-/-}$ compared to their siScrambled transfected control (**G**). Probe intensity values are presented as row-scaled Z-scores. The dendrograms are generated from Spearman correlation values.

associated genes did not simply result from the perturbation of the ESC transcriptional network, we employed $ESC^{Nanog-/-}$, $ESC^{Nanog+/+}$.(201) Before analysing the expression of ESC Super Enhancer associated genes, I confirmed the decreased transcription of the ostensibly

knocked out gene by RT-qPCR²² (figure 5.5A). In the null lines relative to their wild-type controls the target gene was substantially downregulated in expression.

Although both ESRRB and NCOA3 play key roles in ESC self-renewal,(200,301,302,305,306) both ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} populations can maintain self-renewal in culture, albeit to a lesser degree. Figure 5.4B shows the results of a self-renewal assay²³ of the knock out lines and their wild-type controls. In the presence of LIF both ESC^{Esrrb-/-}(left panel, P-value = 0.003) and ESC^{Ncoa3-/-} (right panel, P-value = 0.01) cells show a decrease in the proportions of undifferentiated colonies compared to controls as revealed upon alkaline phosphatase staining. As expected, in the absence of LIF the ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} compared to control populations appear equally differentiated. The reduced colony formation efficiency (figure 5.5C) of the the ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} lines compared to their matching controls may, at least partially, explain the reduction in undifferentiated colonies.

To test the effect of depleting ESRRB on ESC Super Enhancer associated genes, I examined their expression using microarray data pertaining to our *Esrrb* knock out model (unpublished data, shared by Hitoshi Niwa), an *Esrrb* siRNA knock down ESC cell line(307) and an *Ncoa3* siRNA knock down ESC cell line.(301) However, after applying a linear regression model few probes returned a $|\log_2(\text{fold-change})| > 2$. Figure 5.5D shows the distribution of these values and contrasts them with the earliest EpiSC/explanted epiblast time-points vs ESC microarray data published by Kojima *et al.*(5) In fact, only 2 probes did so in all three microarray experiments (figure 5.4E). However, when I examined the expression of key self-renewal/pluripotency and differentiation markers, I found that the former were downregulated and the latter upregulated in the *Esrrb* (figure 5.5F) and *Ncoa3* (figure 5.5G) depleted cells compared to their wild-type counterparts. One would expect to see this phenotype upon successful depletion of these genes. Given the essential roles played by ESRRB and its essential co-activator NCOA3 in ESCs (18,20,248,301,306–308), the impact of their depletion

²² Methods: Chapter 2.2.1.4

²³ Methods: Chapter 2.2.1.2.1

Table 5.4 Genes that appear differentially expressed in the absence of *Esrrb* or *Nanog* according to Rank Products meta-analysis of microarray data.

Rank Products looks for consistently highly ranked items (i.e. genes) in sets of statistics (i.e. *t*-statistics for differential expression upon depletion of *Esrrb* or *Nanog*). The more often a gene ranks highly, the more probable that it truly differs in expression due to the low probability of this occurring by chance. To take into account the directionality of differential expression, *Rank Products* tests each gene for upregulation and downregulation separately. This table lists the number of genes the *Rank Products* evaluated as statistically significantly upregulated (↑) and downregulated (↓) upon depletion of *Esrrb* or *Nanog*, as well as the number of genes not differentially expressed (~).

Expression		Depleted gene	
		<i>Esrrb</i>	<i>Nanog</i>
↑		1431	1099
~		17155	16550
↓		1396	1014

on self-renewal and cell plating efficiency (figures 5.4B and 5.4C), and the differential expression of self-renewal/pluripotency and differentiation markers, I postulated that a meta-analysis of all publicly available microarray experiments²⁴ pertaining to *Esrrb* depletion would give me the power to detect consistent differential mRNA expression events, which I could validate using RT-qPCR. I applied the same methodology to *Nanog* for comparison (For a review of contemporary microarray meta-analysis methodology and statistical considerations see Tseng *et al.*(309)) Briefly, I applied *Rank Products*²⁵ with the *RankProd* Bioconductor package(235), a list ranking-based method of meta-analysis, to find genes that consistently appeared towards the top of the microarray differential expression hits tables and statistically evaluated the likelihood of a given gene occurring at a given position using permutation testing (see figure 6.5A for an overview of the workflow). The Rank Product/Rank Sum (RP/RS) value indicates the position of a given item (i.e. gene) relative to the overall length of the list (i.e. ranked list of *t*-statistics for differential expression). To take into account the directionality of

²⁴ See Appendix Table A.2 for a full list of the microarray experiments used here.

²⁵ Methods: Chapter 2.2.2.3.2

Table 5.5 Rank Products designates *Esrrb* and *Nanog* as statistically significantly downregulated in the *Esrrb* and *Nanog* depleted cell lines compared to their normal ESC controls.

The Rank Product/Rank Sum (RP/RS) value indicates the position of a given item (i.e. gene) relative to the overall length of the list (i.e. ranked list of t-statistics for differential expression). To account for the directionality of differential expression, RankProd tests each gene for upregulation and downregulation separately, giving two sets of output values for each gene. The smaller the RP/RS the more likely that depletion of *Esrrb* or *Nanog* induces differential expression of that gene in that direction. The RankProd package calculates a P-value for differential expression of that gene based on permutation testing. This table shows the RP/RS for the *Esrrb* and *Nanog* genes in the upregulated ($\uparrow$) and downregulated ($\downarrow$) results lists from the *Esrrb* and *Nanog* depleted cell lines compared to their normal ESC controls. Both *Esrrb* and *Nanog* show small RP/RS values in the downregulated gene list for both meta-analyses, indicating that upon depletion of *Esrrb* or *Nanog* both genes lose expression. The depletion target shows the smallest RP/RS in both instances. As expected, *Esrrb* returns a very large RP/RS in the *Esrrb* depleted upregulated gene set, and *Nanog* isn't included in the *Nanog* depleted upregulated gene set, indicating that neither gene increases in expression when targeted for depletion. Put simply when ESCs are depleted of gene X, microarray analysis consistently detects reduced transcription of gene X, and Rank Products correctly reports that gene X has a high probability of downregulation and low probability of upregulation.

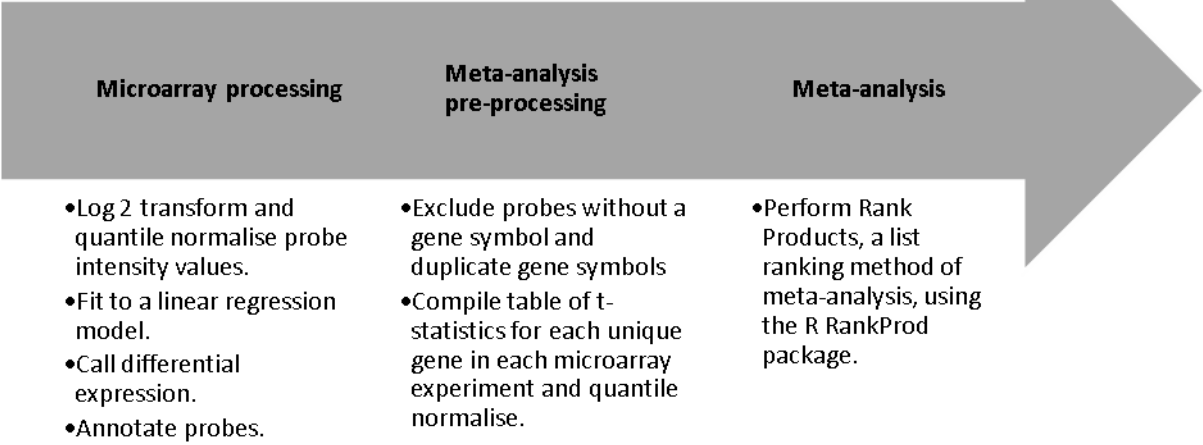
Depleted			Microarray probe			
			<i>Esrrb</i>		<i>Nanog</i>	
			RP/RS	P-value	RP/RS	P-value
	<i>Esrrb</i>	$\downarrow$	28.689	0.0001	419.4559	0.0026
		$\uparrow$	13308.162	0.9744	14624.3309	0.9897
	<i>Nanog</i>	$\downarrow$	1188.893	0.0202	16.4359	0
		$\uparrow$	NA	NA	NA	NA

differential expression, RankProd tests each gene for upregulation and downregulation separately, giving two sets of output values for each gene. For each set of output values, the smaller the RP/RS the more likely that depletion of *Esrrb* or *Nanog* induces differential expression of that gene in that direction. Table 5.4 details the number of genes called statistically significantly differentially expressed, and the number of remaining genes assayed by the array. To test the ability of the meta-analyses to accurately identify differentially expressed genes, I queried the expression of *Esrrb* and *Nanog*. Upon *Esrrb* depletion, *Esrrb* returns a small and statistically significant RP/RS in the Rank Products downregulated results

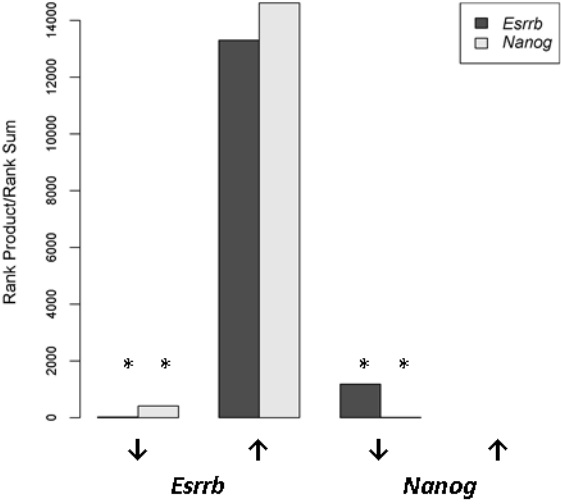
list (RP/RS=29, P-value=0.0001), and a large and non-significant RP/RS in the Rank Products upregulated results list (RP/RS=13308, P-value=0.97) (figure 5.6B, table 5.5). This indicates a high probability for downregulation and low probability of upregulation. *Nanog* returns similar results in the *Nanog* depleted Rank Products downregulated results, and was not present in the upregulated results indicating a very low probability of upregulation (downregulated: RP/RS=16, P-value=0; upregulated: RP/RS=NA, P-value=NA). Therefore, Rank Products can accurately identify differentially expressed genes.

To examine whether Maintained and Silenced Super Enhancer associated genes appear selectively sensitive to the depletion of *ESRRB* and/or *NANOG* I compared their RP/RS values (figure 5.6C). In the *ESRRB* microarray meta-analysis, the Silenced compared to the Maintained Super Enhancer associated genes displayed smaller RP/RS values in the downregulated gene set (P-value = 0.01; figure 5.6C, upper left panel) and larger RP/RS values in the upregulated gene set (P-value = 0.009; figure 5.6C, upper right panel). This indicates that the Silenced compared to the Maintained Super Enhancer associated genes are more likely to lose expression upon *Esrrb* depletion. In contrast, there's no difference in the RP/RS values of Maintained and Silenced associated genes in the *Nanog* microarray meta-analysis. To take into account the Rank Products derived P-values, I used Chi-square testing to see if the Silenced and Maintained ESC Super Enhancer associated genes appear in any list at a greater than expected distribution. Here I found that in the *Esrrb* microarray meta-analysis, the Silenced ESC Super Enhancer associated genes occur less frequently than expected in the upregulated and no change gene lists, and occur at above expected frequency in the downregulated gene list, and conversely, the Maintained ESC Super Enhancer associated genes occur less frequently than expected in the downregulated list (figure 5.6D, left panel; Chi-squared P-value = 0.01). In contrast, there is no difference in the distribution of Silenced and Maintained Super Enhancer associated genes upon *Nanog* depletion (figure 5.6D, right panel; Chi-squared P-value = 0.6). Overall, this further evidences

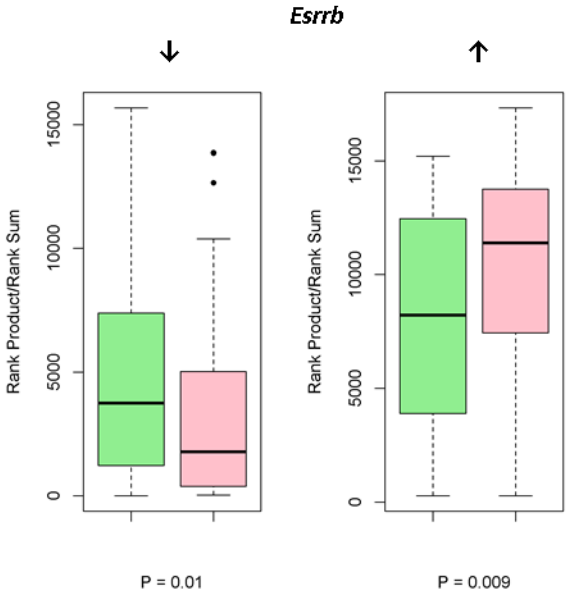
A



B



C



D

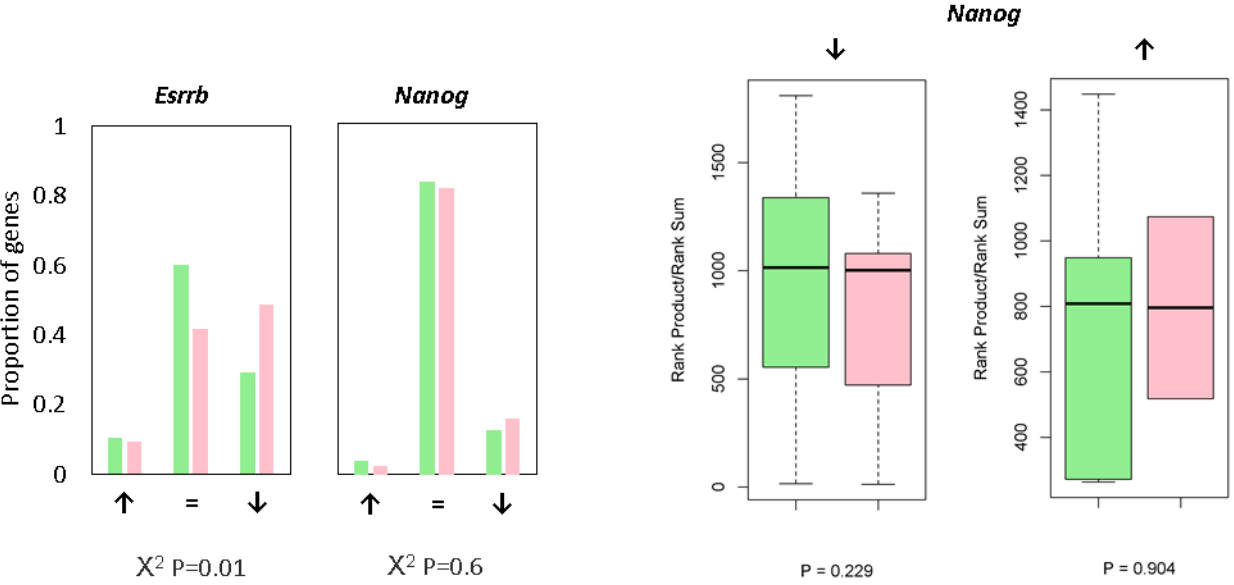


Figure 5.6 Meta-analysis of microarray experiments performed on embryonic stem cells depleted for *Esrrb* and *Nanog*.

A. Schematic representation of the microarray meta-analysis workflow. **B.** Barplot showing the Rank Product/Rank Sum (RP/RS) values of *Esrrb* and *Nanog* in the downregulated and upregulated gene lists outputted by RankProd. RankProd tests each gene for upregulation and downregulation separately, giving two sets of output values for each gene. For each set of output values, the smaller the RP/RS the more likely that depletion of *Esrrb* or *Nanog* induces differential expression of that gene in that direction. The RankProd package calculates a P-value for differential expression of that gene based on permutation testing. This table shows the RP/RS for the *Esrrb* and *Nanog* genes in the upregulated (↑) and downregulated (↓) results lists from the *Esrrb* and *Nanog* depleted cell lines compared to their normal ESC controls. Both *Esrrb* and *Nanog* show small RP/RS values in the downregulated gene list for both meta-analyses, indicating that upon depletion of *Esrrb* or *Nanog* both genes lose expression. The depletion target shows the smallest RP/RS in both instances. As expected, *Esrrb* returns a very large RP/RS in the *Esrrb* depleted upregulated gene set, and *Nanog* isn't included in the *Nanog* depleted upregulated gene set, indicating that neither gene increases in expression when targeted for depletion. The asterisks indicates an FDR corrected P-value < 0.05. Neither *Esrrb* nor *Nanog* were included in the upregulated gene list from the *Nanog* depleted microarray meta-analysis. **C.** Boxplots showing the RP/RS values of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in the upregulated (↑) and downregulated (↓) gene lists outputted by RankProd. The upper panel pertains to the *Esrrb* depleted microarray meta-analysis and the lower panel pertains to the *Nanog* depleted microarray meta-analysis. The P-values are derived from Wilcoxon two-sample t-tests. **D.** Barplots showing the proportion of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in upregulated (↑) and downregulated (↓) gene lists outputted by RankProd. Genes that were present in neither list were designated not differentially expressed (=). The upper panel pertains to the *Esrrb* depleted microarray meta-analysis and the lower panel pertains to the *Nanog* depleted microarray meta-analysis. The P-values are derived from Wilcoxon two-sample t-tests.

a role for *ESRRB* in regulating the activity of Silenced ESC Super Enhancers, while refuting a selective role for *NANOG* at one of these region sets.

5.2.5 *ESRRB* requires its essential co-activator *NCOA3* to mediate its regulation of Silenced Super Enhancers associated genes.

In Chapter 5.2.4, I reported that Silenced compared to Maintained ESC Super Enhancer associated genes appear more sensitive to the depletion of *Esrrb*, while these gene sets showed no differential response to *Nanog* depletion. As discussed in Chapter 5.2.1, microarrays have long been criticised for lack of reproducibility and specificity relative to RT-qPCR. Thus, we validated the results from the microarray meta-analyses by RT-qPCR. We

selected 7 of each the Maintained (*Klf13*, *Lefty1*, *Med13l*, *Pou5f1*, *Otx2*, *Smarcad1*, *Tet1*) and Silenced (*Klf2*, *Klf4*, *Klf5*, *Tbx3*, *Tdh*, *Tet2*, *Tfcp2l1*) Super Enhancer associated genes and examined their expression in the ESC^{Esrrb-/-}, ESC^{Ncoa3-/-}, and ESC^{Nanog-/-} cell lines compared to their wild-type controls. In addition to myself, King Hang Mau performed cell culture, RNA extraction, cDNA synthesis, and RT-qPCR for these experiments. I analysed and visualised all of the resulting data. Firstly, to confirm the fidelity of these cell lines we examined the expression of the deleted genes (figure 5.6A). In each of the null relative to wild-type control cell lines, the deleted gene showed a dramatic downregulation of expression (Two-tailed, unequal variance T-test P-values: *Esrrb* in ESC^{Esrrb-/-} = 0.003, *Ncoa3* in ESC^{Ncoa3-/-} = 0.007, and *Nanog* in ESC^{Nanog-/-} = 0.004). In contrast, the non-deleted genes show no change in transcription.

We next examined the expression of the candidate Maintained and Silenced Super Enhancer associated genes in these cell lines. Figure 5.7B and table 5.6 present the log₂(fold-change) in expression of these genes between the null and wild-type control cell lines. The dendrograms of the cell lines and candidate genes use Spearman correlation co-efficients to represent distance. In all three cell lines the majority of differential expression events occur as a downregulation. The two early post-implantation epiblast markers *Lefty1* and *Otx2* present two exceptions, showing upregulation in both the ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} cell lines, and in the case of *Otx2* in the ESC^{Nanog-/-} cell line also. Interestingly, *Otx2* and *Nanog* have been shown to be expressed in a mutually exclusive manner in ESCs.(310) The largest changes in expression appear to occur in the ESC^{Esrrb-/-} followed by the ESC^{Ncoa3-/-} cell line. Accordingly, the cell line dendrogram clusters together these two ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} cell lines ($\rho = 0.70$, P-value = 0.005), reflecting their greater similarity in the differential expression of target genes compared to the ESC^{Nanog-/-} cell line (ESC^{Esrrb-/-} vs ESC^{Nanog-/-}: $\rho = 0.15$, P-value = 0.61, ESC^{Ncoa3-/-} vs ESC^{Nanog-/-}: $\rho = 0.15$, P-value = 0.60). This observation is unsurprising as *Ncoa3* functions as the co-activator of *Esrrb* in ESCs.(301) For the target genes, the first two clusters broadly segregate the Maintained and Silenced Super Enhancer associated genes, with the

exception of the Maintained Super Enhancer associated genes *Pou5f1* and *Smarcad1*, which cluster with the Silenced Super Enhancer associated genes, which consistently show lower levels of expression in the three null lines.. This indicates that a higher or more stable expression pattern of these genes may associate with the naïve rather than the primed pluripotent state as recently reported.(311,312)

To further assess the impact of *Esrrb*, *Ncoa3*, and *Nanog* loss on the candidate Maintained and Silenced ESC Super Enhancer associated genes, I visualised the $\log_2(\text{fold-changes})$ for each cell line as a waterfall plot, which ranks the candidate genes by increasing differential expression (figure 5.7C). For both ESC^{Esrrb-/-} and ESC^{Ncoa3-/-} cell lines, Silenced Super Enhancer associated genes tend to show the biggest negative differential expression, while Maintained Super Enhancer associated genes tend to show the biggest positive differential expression. In contrast, ESCs^{Nanog-/-} show no clear segregation, thus reiterating that Nanog is not be a major differential regulator of these two sets of genes. Figure 5.7D shows a side-by-side comparison of differential expression of the Maintained and Silenced Super Enhancer associated genes in the three cell lines. Overall, we here confirm that the Silenced Super Enhancer associated genes selectively lose expression in the absence of *Esrrb* and *Ncoa3* (ESC^{Esrrb-/-}: median $\log_2(\text{fold-change})$ – Maintained = 0.1, Silenced = -1.57; P-value = 0.008. ESC^{Ncoa3-/-}: median $\log_2(\text{fold-change})$ – Maintained = -0.1, Silenced = -0.53; P-value = 0.136). Although the differences in Maintained and Silenced Super Enhancer genes in ESC^{Ncoa3-/-} are not statistically significant, the P-value approaches significant and thus this may be due to insufficient power. A larger pool of candidate genes may have shown a statistically significant difference. Alternatively, *Ncoa3* depletion measures the effect of ESRRB inactivation indirectly through inhibiting a co-activator rather than the transcriptional activator itself. Regardless, the high correlation with the ESC^{Esrrb-/-} suggests that the trend is replicated

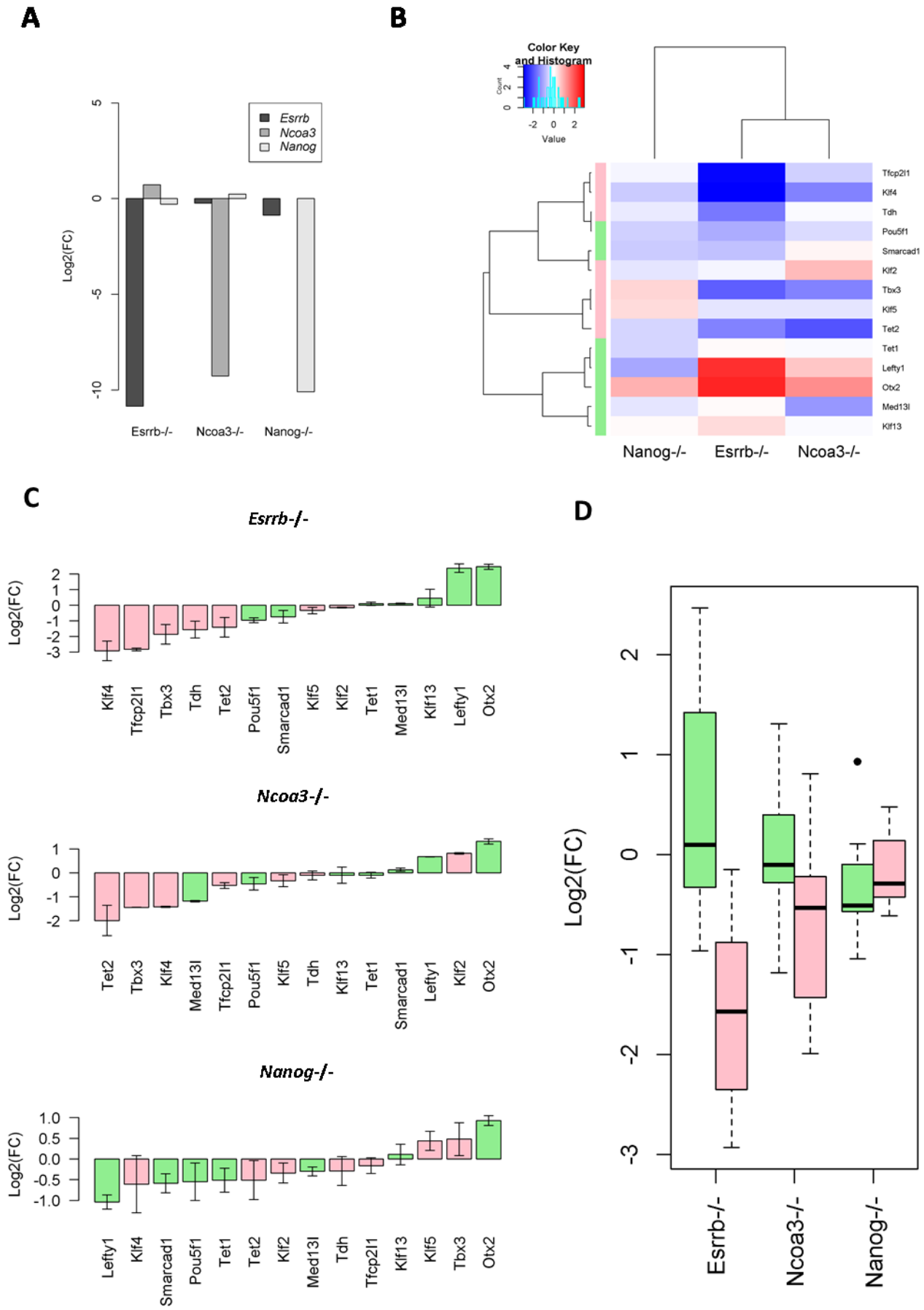


Figure 5.7 Comparative RT-qPCR analysis of Maintained and Silenced Super Enhancer associated gene expression in embryonic stem cells constitutively depleted for Esrrb, Ncoa3 or Nanog.

A. Barplots showing the $\log_2(\text{fold-change})$ in expression of Esrrb, Ncoa3, and Nanog in the ESC lines with homozygous deletions of these genes compared to their wild-type controls. Data were normalised to the housekeeping genes S17 and L19 and expressed as $\log_2(2^{-\Delta\Delta CT})$. Three passages were collected per cell line and each qPCR was performed in triplicate. The error bars represent standard error from the mean (SEM). **B.** Heatmaps showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-}, ESC^{Ncoa3-/-} and ESC^{Nanog-/-} lines compared to their respective controls. Probe intensity values are presented as row-scaled Z-scores. The dendrograms are generated from Spearman correlation values. **C.** Waterfall plots showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-}, ESC^{Ncoa3-/-} and ESC^{Nanog-/-} lines compared to their respective controls. As in part A, the error bars represent the SEM of the biological replicates. **D.** Boxplots showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-}, ESC^{Ncoa3-/-} and ESC^{Nanog-/-} lines compared to their respective controls.

Table 5.6 The expression of Maintained and Silenced Super Enhancer associated genes in ESCs bearing homozygous deletions of Esrrb, Ncoa3 or Nanog compared to their wild-type controls.

Data were normalised to the housekeeping genes S17 and L19 and expressed as $\log_2(2^{-\Delta\Delta CT})$, listed as $\log_2(\text{FC})$ in this table. Three passages were collected per cell line and each qPCR was performed in triplicate. The P-values derive from unpaired, unequal variance T-tests. Target – the transcript amplified by qPCR; ESC SE Class – the class of ESC Super Enhancer (SE) the target belongs to. The red-blue colour scheme indicates whether a target is upregulated (red) or downregulated (blue). The cells highlighted in green in the P-value columns indicate $P < 0.05$.

Target	ESC SE Class	ESC ^{Esrrb-/-}		ESC ^{Ncoa3-/-}		ESC ^{Nanog-/-}	
		Log ₂ (FC)	P-value	Log ₂ (FC)	P-value	Log ₂ (FC)	P-value
Klf13	Maintained	0.46	0.30	-0.10	0.58	0.11	0.80
Lefty1	Maintained	2.38	0.01	0.67	0.39	-1.04	0.13
Med13l	Maintained	0.10	0.67	-1.18	0.00	-0.30	0.55
Otx2	Maintained	2.47	0.02	1.31	0.02	0.93	0.24
Pou5f1	Maintained	-0.96	0.05	-0.46	0.09	-0.55	0.25
Smardc1	Maintained	-0.74	0.03	0.13	0.64	-0.59	0.08
Tet1	Maintained	0.09	0.54	-0.10	0.64	-0.51	0.11
Klf2	Silenced	-0.15	0.56	0.81	0.37	-0.34	0.27
Klf4	Silenced	-2.93	0.00	-1.42	0.04	-0.61	0.28
Klf5	Silenced	-0.34	0.41	-0.33	0.03	0.44	0.22
Tbx3	Silenced	-1.87	0.15	-1.44	0.03	0.48	0.63
Tdh	Silenced	-1.57	0.23	-0.11	0.33	-0.29	0.12
Tet2	Silenced	-1.41	0.04	-1.99	0.01	-0.51	0.37
Tfcp2l1	Silenced	-2.83	0.07	-0.53	0.36	-0.16	0.63

in the ESC^{Ncoa3^{-/-}} cell line. The Maintained Super Enhancer associated genes show no evident change in their relative expression in ESC^{Nanog^{-/-}}, apart from a slight and non-significant decrease in the expression of the Maintained Super Enhancer associated genes (ESC^{Nanog^{-/-}}: median log₂(fold-change) – Maintained = -0.51, Silenced = -0.29; P-value = 0.647).

To further evidence that *Esrrb* depletion directly causes the downregulation of Silenced ESC Super Enhancer associated genes in the ESC^{Esrrb^{-/-}} cell line, we performed a rescue experiment. We transfected ESCs with either an empty CAG vector, wild-type *Esrrb*, or *Esrrb* with a non-functional AF2 domain interfering with NCOA3 binding to ESRRB and, in turn, ESRRB transcriptional activity.(301) Figure 5.8 summarises the stable generation of these cell lines. We then assayed the expression of the same candidate Maintained and Silenced ESC Super Enhancer associated gene panel in all three cell lines. Dr Rute Tomaz transfected the cells and harvested cell pellets. King Hang Mau isolated RNA, synthesised cDNA, and performed the RT-qPCR. I analysed and visualised the resulting data. I normalised the ΔC_T values of the ESC^{Esrrb^{-/-}}|*Esrrb* and ESC^{Esrrb^{-/-}}|*Esrrb*ΔAF2 lines to those of the ESC^{Esrrb^{-/-}}|Empty line. Figure 5.9A and table 6.8 show the log₂(fold-change) in expression of the two *Esrrb* expressing lines compared to ESC^{Esrrb^{-/-}}|Empty we confirmed the rescued expression of wild-type and mutant

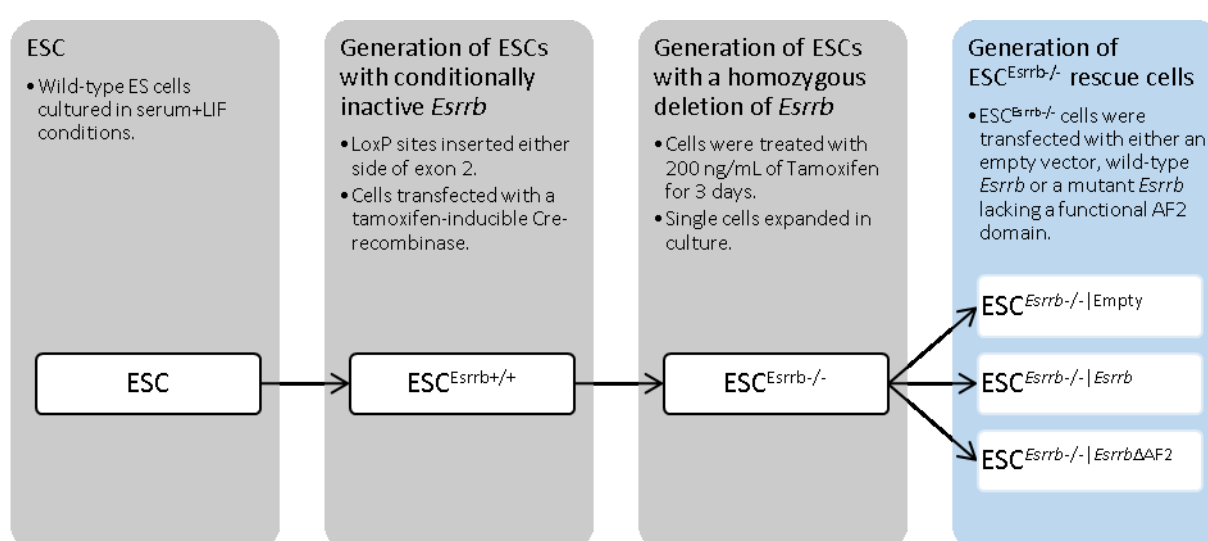


Figure 5.8 The generation of ESC^{Esrrb^{-/-}} rescue cells.

Esrrb forms (ESC^{*Esrrb*^{-/-}|*Esrrb*} $\log_2(\text{FC}) = 15.84$, P-value = 0.012. ESC^{*Esrrb*^{-/-}|*Esrrb* ΔAF2} $\log_2(\text{FC}) = 16.54$, P-value = 0.002). To overview the changes in gene expression upon transfection and puromycin selection, I first plotted the $\log_2(\text{fold-changes})$ as a heatmap. I included the differential expression values from the non-transfected ESC^{*Esrrb*^{-/-}} line normalised to ESC^{*Esrrb*^{+/+}} control for comparison as shown in figure 5.7D. The reintroduction of wild-type *Esrrb* appears to induce expression of almost all the candidate ESC Super Enhancer associated genes. However, with the exception of Silenced *Klf2* ($\log_2(\text{FC}) = 0.83$, P-value = 0.08) and Maintained *Klf13* ($\log_2(\text{FC}) = 0.87$, P-value = 0.01), only the differential expression of the Silenced Super Enhancer associated genes prove statistically significant, supportive of *Esrrb*-mediated regulation of these genes. Interestingly, the transfection of the mutant *Esrrb* largely downregulates both Maintained and Silenced ESC Super Enhancer associated genes. . ESRRB requires its co-activator NCOA3 to drive gene expression. NCOA3 interacts with ESRRB via the ESRRB AF-2 domain. It's possible that the downregulation of both groups of ESC Super Enhancer associated genes could be because mutating the AF-2 domain results in a dominant negative ESRRB. Its continued binding to regulatory DNA regions prevents the cell from establishing the compensation mechanisms we see in the ESC^{*Esrrb*^{-/-}} cells or destabilise the network of proteins at enhancers. If so, one might expect to see ESC^{*Esrrb*^{-/-}|*Esrrb* ΔAF2} compared to ESC^{*Esrrb*^{-/-}|Empty} and ESC^{*Esrrb*^{-/-}} cells exhibit a further decrease in colony formation efficiency and self-renewal capacity although this has not been formally tested in this study.

To globally assess whether the *Esrrb* rescues selectively impact on the candidate Maintained or Silenced ESC Super Enhancer associated genes, I replicated the waterfall plots from figure 5.7C (figure 5.9C). In ESCs^{*Esrrb*^{-/-}|*Esrrb*}, Silenced Super Enhancer show the greatest increases in differential expression. In ESCs^{*Esrrb*^{-/-}|*Esrrb* ΔAF2} , neither group of genes show a clear trend. Figure 5.9D summarises these results as boxplots. In ESCs^{*Esrrb*^{-/-}|*Esrrb*} the Silenced compared to the Maintained ESC Super Enhancer associated genes show a substantial increase in

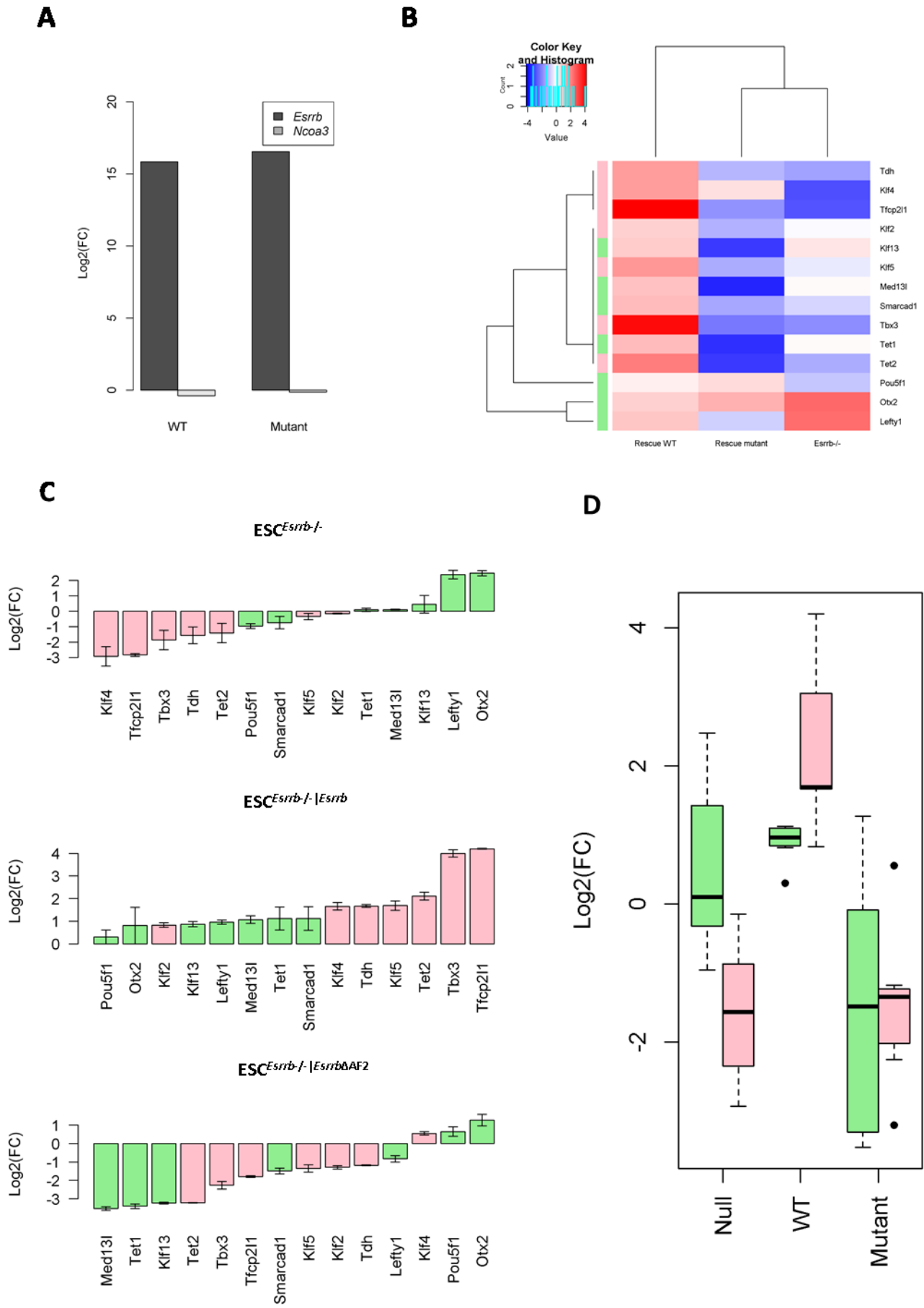


Figure 5.9 ESRRB requires a functional AF2 domain to positively regulate the expression of Silenced Super Enhancer associated genes in ESCs.

A. Barplots showing the $\log_2(\text{fold-change})$ in expression of *Esrrb* and *Ncoa3* in ESC lines with homozygous deletion of *Esrrb* rescued with wild-type *Esrrb* and *Esrrb* Δ AF-2 compared to those transfected with an empty vector. Data were normalised to the housekeeping genes *S17* and *L19* and expressed as $\log_2(2^{-\Delta\Delta CT})$. Three passages were collected per cell line and each qPCR was performed in triplicate. The error bars represent standard error from the mean (SEM). **B.** Heatmaps showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-|Esrrb}, ESC^{Esrrb-/-|Esrrb} Δ AF-2 compared to ESC^{Esrrb-/-|Empty}, and ESC^{Esrrb-/-} compared to wild-type ESCs. Normalised probe intensity values are presented as row-scaled Z-scores. The dendrograms are generated from Spearman correlation values. **C.** Waterfall plots showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-|Esrrb}, ESC^{Esrrb-/-|Esrrb} Δ AF-2 compared to ESC^{Esrrb-/-|Empty}, and ESC^{Esrrb-/-} compared to wild-type ESCs. The error bars represent standard error. **D.** Boxplots showing the $\log_2(\text{fold-change})$ in expression of Maintained (light green) and Silenced (pink) ESC Super Enhancer associated genes in ESC^{Esrrb-/-|Esrrb}, ESC^{Esrrb-/-|Esrrb} Δ AF-2 compared to ESC^{Esrrb-/-|Empty}, and ESC^{Esrrb-/-} compared to wild-type ESCs.

expression (median $\log_2(\text{FC})$: Maintained = 0.96, Silenced = 1.69. P-value = 0.026). In contrast, ESCs^{Esrrb-/-|Esrrb} Δ AF2 show no difference in the expression pattern of these two groups of genes. In both figure 5.9C and 5.9D, results obtained from non-transfected ESCs^{Esrrb-/-} normalised to ESCs^{Esrrb+/+} are shown for comparison.

In conclusion, the RT-qPCR data confirms that Silenced compared to Maintained ESC Super Enhancer associated genes appear more sensitive to *Esrrb* expression and transcriptional activity. This functionally evidences a role for ESRRB and NCOA3 in regulating Silenced ESC Super Enhancers via DM Super Enhancelets.

5.2.6 ESRRB loss selectively depletes MED1 at DM Super Enhancelets

Whyte *et al.* initially defined ESC Super Enhancers on the basis of high MED1 binding density across large genomic regions. Hnisz *et al.* further reported that one can use H3K27ac in the place of MED1 as a universal Super Enhancer defining factor. (195) If ESRRB preferentially regulates DM as opposed to PU ESC Super Enhancelets, one might expect to see a region-specific loss of H3K27ac and/or MED1. To test this hypothesis, we performed ChIP-seq for

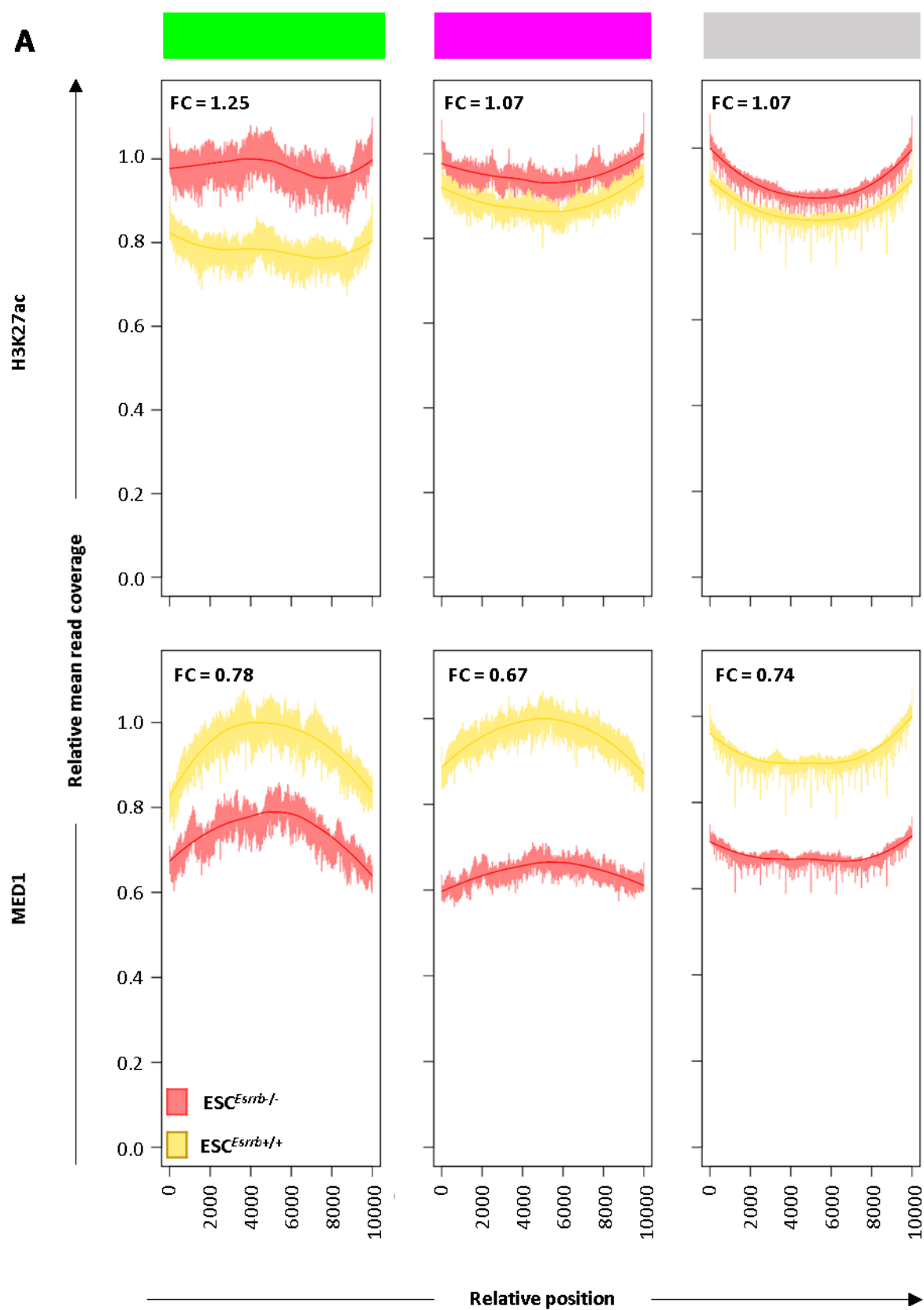
H3K27ac and MED1 in ESC^{Esrrb-/-} and ESC^{Esrrb+/+} lines. Dr Rute Tomaz cultured these cell lines, extracted the chromatin, performed the ChIP, and prepared the sequencing libraries. Samples were sequenced at the Radboud Institute for Molecular Life Sciences in collaboration with Professor Henk Stunnenberg. I analysed the resulting data. Figure 5.10A shows the relative mean read coverage of H3K27ac and MED1 ChIP-seq data at PU, DM and PM subregions. Figure 5.10B shows the mean fold-change for each of these factors and subregions between the null and wild-type cell lines.

Unexpectedly, although similarly distributed at DM and PM subregions, H3K27ac appears enriched at PU subregions (mean FC = 1.25). If ESRRB forms part of an independent group of proteins that drive the activity, it's presence at PU may act in a competitive manner, preventing PU-specific transcription factors from interacting with DNA, therefore exerting a

Table 5.7 The expression of Maintained and Silenced Super Enhancer associated genes in ESCs bearing a homozygous deletion of Esrrb, rescued with either wild-type Esrrb or mutant Esrrb, compared to their wild-type controls.

Data were normalised to the housekeeping genes S17 and L19 and expressed as $\log_2(2^{-\Delta\Delta CT})$, listed as $\log_2(FC)$ in this table. Three passages were collected per cell line and each qPCR was performed in triplicate. The P-values derive from unpaired, unequal variance T-tests. Target – the transcript amplified by qPCR; ESC SE Class – the class of ESC Super Enhancer (SE) the target belongs to. The red-blue colour scheme indicates whether a target is upregulated (red) or downregulated (blue). The cells highlighted in green in the P-value columns indicate $P < 0.05$.

Target	ESC SE Class	ESC ^{Esrrb-/-} Esrrb		ESC ^{Esrrb-/-} Esrrb Δ AF-2	
		Log ₂ (FC)	P-value	Log ₂ (FC)	P-value
<i>Klf13</i>	Maintained	0.87	0.01	-3.22	0.00
<i>Lefty1</i>	Maintained	0.96	0.15	-0.83	0.21
<i>Med13l</i>	Maintained	1.07	0.07	-3.53	0.01
<i>Otx2</i>	Maintained	0.81	0.25	1.27	0.25
<i>Pou5f1</i>	Maintained	0.30	0.28	0.65	0.07
<i>Smarcad1</i>	Maintained	1.12	0.15	-1.49	0.08
<i>Tet1</i>	Maintained	1.12	0.10	-3.40	0.01
<i>Klf2</i>	Silenced	0.83	0.08	-1.29	0.04
<i>Klf4</i>	Silenced	1.66	0.00	0.55	0.06
<i>Klf5</i>	Silenced	1.69	0.00	-1.35	0.00
<i>Tbx3</i>	Silenced	3.99	0.00	-2.26	0.00
<i>Tdh</i>	Silenced	1.67	0.01	-1.18	0.04
<i>Tet2</i>	Silenced	2.11	0.00	-3.21	0.00
<i>Tfcp2l1</i>	Silenced	4.20	0.00	-1.79	0.01



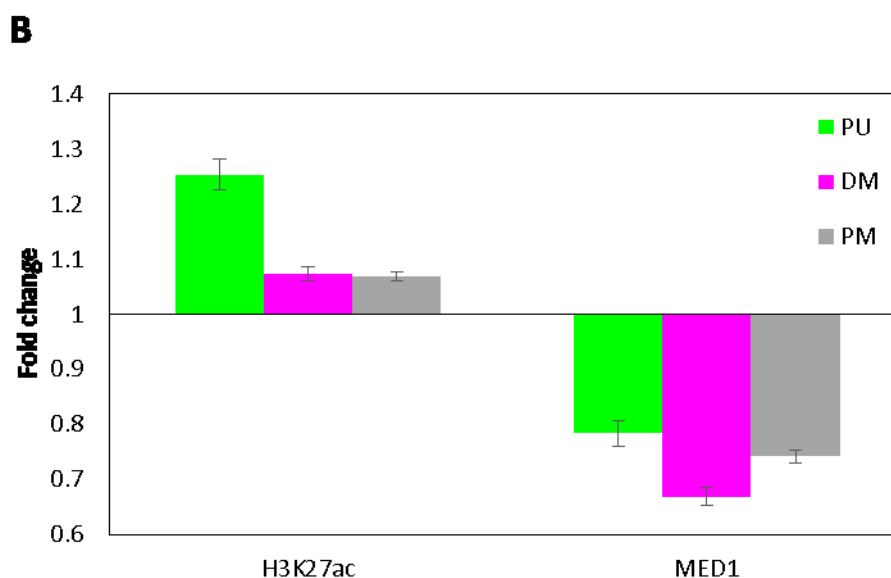


Figure 5.10 Assessing the impact of *ESRRB* loss on H3K27a deposition and MED1 binding at DM and PU Super Enhancers in ESC.

A. Metagenes showing the mean ChIP-seq read coverage of H3K27ac and MED1 at PU, DM and PM ESC Super Enhancer subregions in $ESC^{Esrrb-/-}$ (red) and $ESC^{Esrrb+/+}$ (yellow). ChIP-seq bam files were normalised for sequencing depth and input. Mean read coverage values were calculated using Dr Ed Curry's plotCoverage function (see Method section), and scaled to the maximal value at each region set. **B.** The mean fold-change in ChIP-seq read coverage of H3K27ac and MED1 in $ESC^{Esrrb-/-}$ compared to $ESC^{Esrrb+/+}$. The error bars represent the standard deviation

minor inhibitive effect. Alternatively, if the *ESRRB*-containing protein subgroup functions to prepare DM for decommissioning, the same minor inhibitive effect might be observed. In either case the loss of *ESRRB* might bolster the PU enhancer complexes. This could at least partially explain why some Maintained ESC Super Enhancer associated genes increase in expression in the absence of *Esrrb* (figure 5.7). In contrast to H3K27ac, MED1 shows a consistent depletion at all three subregions, however DM show the greatest loss (mean FC = 0.67). This suggests that in the absence of *ESRRB*, MED1 binding to Super Enhancers destabilises globally, but DM appear the most heavily affected. Protein-protein interaction data shows that

ESRRB physically associates with multiple members of the Mediator complex, including MED1, independently from OCT4.(249) Thus ESRRB may assist in recruiting MED1 to ESC Super Enhancers, and DM rely on ESRRB to do so more than PU.

5.3 Conclusion

In this chapter I aimed to identify transcription factors that differentially regulate PU and DM ESC Super Enhancelets. My first step was to use a publicly available ESC shRNA screen (figure 5.1). This highlighted the naïve pluripotency factors ESRRB, NCOA3, KLF5, and NANOG as factors upon whose depletion resulted in a greater loss of Silenced compared to Maintained Super Enhancer associated genes, while depletion of the primed pluripotency factors SOX2 and OTX2 led to a greater loss of Maintained compared to Silenced Super Enhancer associated genes. I posited that the DNA binding motifs of candidate differential regulators might enrich at one set of Super Enhancelets over the other. This identified the ESRRB and KLF5 binding motifs as specifically enriched at DM compared to PU, and the CTCF and, surprisingly, NANOG binding motifs at PU compared to DM (figure 5.2). I then looked for differences in patterns of protein binding density at PU and DM subregions. Relative to all other candidate differential regulators, ESRRB and NCOA3 uniquely showed similar binding coverage at DM and PU. Given that PU exhibits enhanced chromatin accessibility compared to DM, this might indicate a selective and high level of enrichment for ESRRB and NCOA3 binding at DM as further revealed by recalculating read coverage relative to Med1. To confirm that the depletion of *Esrrb*/*Ncoa3* preferentially impacted on Maintained and/or Silenced Super Enhancer associated gene expression, I conducted a meta-analysis of all publicly available microarray data generated in ESCs upon depletion of either *Esrrb* or *Ncoa3*, as well as *Nanog* for comparison. Here, *Esrrb* and *Ncoa3* depletion selective reduced expression of Silenced Super Enhancer associated genes implying a prime role for these molecules in regulating DM Super Enhancelets. These results were further supported by RT-qPCR analysis of candidate Maintained and Silenced Super Enhancer associated genes in ESCs bearing homozygous deletions of *Esrrb* and *Ncoa3*, which validated the results of the microarray meta-analysis. Furthermore, the same RT-qPCR analysis in ESCs bearing a homozygous deletion of *Esrrb* and rescued with wild-type or mutant *Esrrb* evidenced a direct role for ESRRB and NCOA3 in regulating Silenced SE associated gene expression in rescue

experiments. Finally, to test the stability of the enhancer machinery at these subregions in the absence of ESRRB, I looked at MED1 binding density assayed by ChIP-seq. Compared to PU, DM showed a greater loss in MED1 binding. These results are summarised in table 5.8. Collectively, these findings demonstrate that ESRRB is a limiting factor in the activity of DM, but not PU ESC Super Enhancelets.

Table 5.8 Candidate ESC Super Enhancelet differential regulators.

Protein – the candidate regulator; *SElet* – the Super Enhancelet they may regulate; *Line of evidence* – whether the an experiment supported (●) or refuted (○) a given candidate regulating a set of Super Enhancelets. NA indicates that the candidate was removed from or not relevant to an experiment.

Protein	SElet	Line of evidence					
		shRNA screen	Motif enrichment	Protein binding density	Microarray meta-analysis	RT-qPCR	MED1 ChIP-seq
ESRRB	DM	●	●	●	●	●	●
NCOA3	DM	●	NA	●	●	●	NA
CTCF	PU	○	●	○	NA	NA	NA
KLF5	DM	●	●	○	NA	NA	NA
NANOG	DM	●	○	○	○	○	NA
NANOG	PU	○	●	○	○	○	NA
SOX2	PU	●	○	○	NA	NA	NA
OTX2	PU	●	○	○	NA	NA	NA

We are much too much inclined in these days to divide people²⁶ into permanent categories, forgetting that a category only exists for its special purpose and must be forgotten as soon as that purpose is served.

Dorothy L. Sayers, 'Are Women Human?'

²⁶ Read: genomic regulatory elements.

6 DISCUSSION

6.1 The signal: ESC Super Enhancers and Typical Enhancers during pluripotency maturation

This thesis grew from Prof Richard Young's group's 2013 triad of publications characterising a set of tissue-specific enhancers that they dubbed Super Enhancers.(179,194,195) They defined Super Enhancers in mouse ESCs as enhancers containing high density MED1 binding. Low density MED1 binding sites they termed Typical Enhancers. Super Enhancers comprise multiple constituent enhancers spread over wide genomic regions (median = 10 KB). Key ESC transcription factors bind to Super Enhancers at high density. Compared to Typical Enhancers, the genes associated with Super Enhancers show markedly greater transcription. At the time we did not know whether ESC Super Enhancers restricted their activity to the naïve pluripotency network of the ESCs in which they were defined, or whether they continued to drive transcription under the primed pluripotency of EpiSCs. Given that the ESC transcriptional network represents one of the best characterised, if the former were the case, ESC Super Enhancers during pluripotency maturation could provide a model to study how tissue-specific enhancers decommission. With this motivation in mind I proceeded to determine how the epigenetic landscape of ESC Super Enhancers and typical enhancers is remodelled in EpiSCs primarily using CpG methylation as a proxy for enhancer activity.

6.1.1 Do ESC Super Enhancers and Typical Enhancers retain the same open epigenetic signature in EpiSCs?

CpG methylation of enhancers correlates with chromatin condensation and transcriptional repression.(74–80) Previous work published by Wiench *et al.*, demonstrates that one can use CpG methylation to identify cell-type specific as opposed to constitutively active enhancers, and distinguish the active from the inactive.(77) If Super Enhancers and Typical Enhancers respectively act in a cell-type specific and constitutively active manner, and ESC Super Enhancers represent naïve pluripotency, one might expect to see that ESC Super Enhancers

selectively acquire CpG methylation in EpiSCs. The differing sequence composition of ESC Super Enhancers and Typical Enhancers supported this premise. ESC Super Enhancers compared to Typical Enhancers harbour a greater GC% and normalised CpG fraction (figure 3.1B). Given the increased mutability of 5-mC, (41–43,48,49) this implies stronger evolutionarily selective pressure to preserve the CpGs at Super Enhancers. I propose two possible explanations: either CpG methylation more actively regulates Super Enhancer function necessitating their retention, or this reflects a general increase in sequence conservation of Super Enhancers perhaps related to preserving transcription factor binding sites. Although difficult to disprove the latter, comparing the sequence conservation of the CpGs to the surrounding nucleotides would show whether they're subject to multiplicative selective pressure effects. I predicted that Super Enhancers and Typical Enhancers would appear hypomethylated in ESCs, and that Typical Enhancers would remain so in EpiSCs, while Super Enhancers would either also remain hypomethylated or acquire methylation, depending on their cell-type specificity. We found that Super Enhancers harbour relatively low CpG methylation in ESCs and this dramatically increases in EpiSCs (figure 3.1C). This evidences that Super Enhancers may decommission in EpiSCs. However, the Typical Enhancers CpG methylation profile deviated from my predictions. Approximately half of Typical Enhancers befall the same fate as Super Enhancers. The remaining half appear highly CpG methylated in ESCs as well as in EpiSCs. This implies that methylated half of the Typical Enhancers defined by Whyte *et al.*, do not actively promote transcription in ESCs, and unmethylated half decommission in EpiSCs. Re-organisation of the epigenome accompanies the shift in transcriptome and cell signalling pathway dependency during pluripotency maturation.(5,10,11,129,130) CpG methylation gradually increases from the late pre-implantation blastocyst stage,(123,124) specifically in the pluripotent ICM relative to the extra-embryonic TE.(124) Thus the increase in CpG methylation at Super Enhancers and Typical Enhancers may simply reflect a global trend. However, ProB cell Super Enhancers – presumably dormant in ESCs – show similarly high levels of CpG methylation to their 5' and

3' flanking regions in ESCs (figure 3.2C). This suggests ESCs actively maintain CpG hypomethylation at ESC Super Enhancers.

Ultimately, the functional output of enhancer activity is gene expression. CpG methylation of gene promoters strongly negatively correlates with expression.(67–73) Therefore, if the ESC Super Enhancers decommission in EpiSCs, one might expect to see a concurrent increase in CpG methylation of their associated genes promoters. However, identifying the genes that enhancers regulate is challenging. Whyte *et al.*, associated Super Enhancers with genes by determining the most highly expressed gene within 50 KB. They validated these interactions by comparison to a topological domains, defined by HiC-seq,(181) and Enhancer-Promoter Unit (EPU) assignments defined by integrated HiC-seq, histone modification and RNA polymerase II ChIP-seq data.(313) Although simple, this approach presents several shortcomings: higher order chromatin folding enables enhancers to interact with promoters up to a megabase away(190,191,314); multiple enhancers interact with the same promoters and other genomic regions and *vice versa* both in *cis* and in *trans* (189,190); and 4C-seq experiments in *Drosophila* indicate that enhancer-promoter loops pre-form early in development and arrest in a transcriptionally paused position until they become relevant to the cell, thus a loop doesn't necessarily reflect an active enhancer-promoter relationship.(189) In short, detecting active promoter-enhancer interactions requires an integrated approach bringing together information about nucleosome occupancy, transcription factor binding density, histone modification profiles, CpG methylation profiles, DNA looping events, RNA polymerase II activity, enhancer RNA (eRNA) expression, and mRNA expression of the candidate gene. Some of the experiments required to assay these features are by no means inexpensive, nor simple to analyse, however the ever-increasing wealth of publicly available sequencing data makes this goal more achievable by the day. In fact, in June of this year, Feng *et al.*, published a tool that does just this for several human cell lines.(315)

Regardless of the shortcomings of their approach, Whyte *et al.*, did identify enhancer-promoter interactions previously validated by 3C-qPCR.(316) Analysing HiC-seq data from ESCs

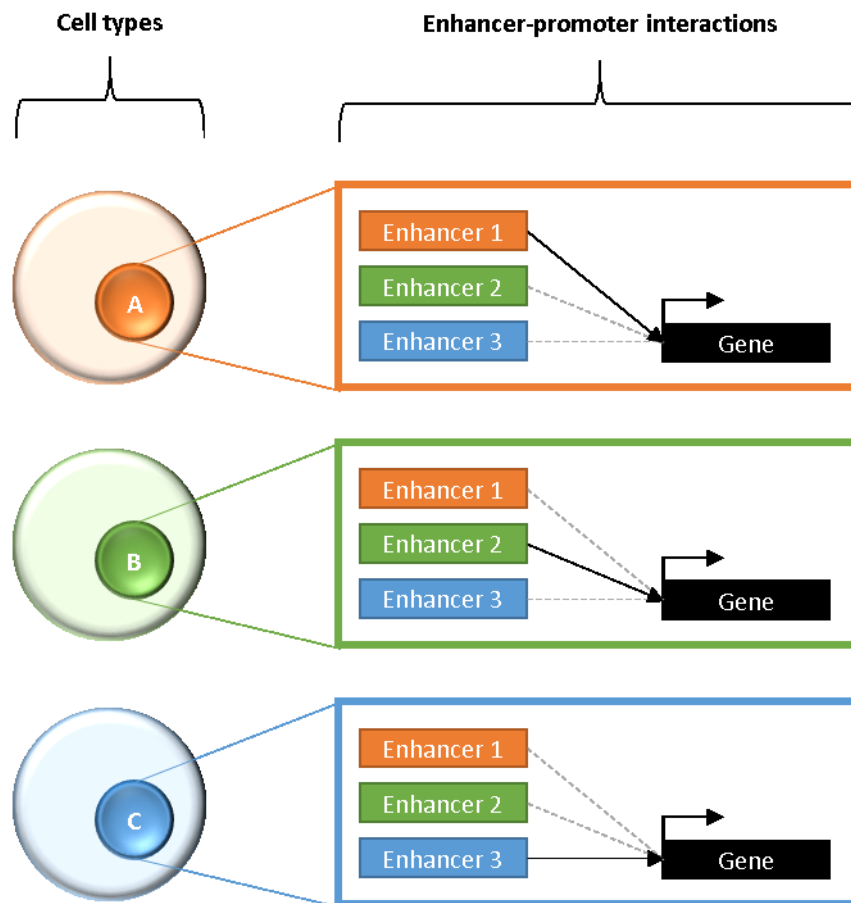


Figure 6.1 Epigenetic remodelling of enhancers rather than promoters allows multiple enhancers to drive the same gene under different transcriptional programmes.

*Enhancer-promoter interactions form early in embryonic development and lie dormant until the appropriate cellular environment arises. Developmental genes play multiple roles throughout development. For example, *Esrrb* forms part of both the naive ESC and TS cell transcriptional network. If epigenetic remodelling occurs primarily at enhancers rather than promoters, this would allow promoters to maintain a relatively open chromatin state, ready to initiate large-scale transcription when the appropriate cellular context arises and their associated enhancer activates.*

(serum+LIF), we found that of the detectable promoter-enhancer interactions, 65% involved the same enhancer and promoter pair assigned by Whyte *et al.* (figure 3.4A). With the knowledge that an unknown proportion of these assigned genes present incorrect enhancer-promoter assignments, and reduce enhancer-promoter interaction to an overly simplistic one-to-one relationship, we thus concluded that this list of associated genes sufficiently reflects

accurate, active enhancer-promoter interactions to detect changes in promoter CpG methylation during pluripotency maturation. As predicted, we found that promoter CpG methylation tends to increase in tandem with that of Super Enhancers during pluripotency maturation (figure 3.5A). However, this increase appears much slighter. One could argue that this suggests greater epigenetic remodelling events occur at enhancers relative to promoters. As some ESC Super Enhancer associated genes (e.g. *Esrrb*, *Nanog*, *Klf4*) become re-expressed later during development, this would enable the promoters to remain in a relatively open chromatin state, ready to re-initiate large-scale transcription when the appropriate cellular context arises (see figure 6.1 for a schematic). This however assumes an equivalent response to the same magnitude of change (i.e. chromatin responds similarly to an X% increase in CpG methylation at promoters and enhancers).

CpG methylation can both positively and negatively impact on transcription factor binding.(102) Some transcription factors display highly context specific responses to CpG methylation. For example, the key naïve pluripotency transcription factor KLF4 shows a stronger and weaker association with two different versions of its DNA binding motif when the CpG it contains is methylated.(103,104) Some whole families of transcription factors display consistent responses to CpG methylation. For example, ~50% of nuclear receptors, ~60% of homeodomain proteins, and ~80% of POU proteins, including OCT4, increase in binding in response to CpG methylation of variants of their DNA binding motifs, whereas 100% of AP2 proteins, ~40% of bHLH proteins, and ~60% of bZIP proteins decrease in binding.(102) Thus the increase in CpG methylation at ESC Super Enhancers during pluripotency maturation could alter the affinity of transcription factors to their DNA binding motifs. Furthermore, as CpG methylation strongly correlates with chromatin condensation, the increased CpG methylation could occlude binding via reduced access to DNA binding motifs. The lower density of protein binding at hyper- relative to hypomethylated Typical Enhancers evidences this relationship (figure 3.1D).

In summary, ESC Super Enhancers and Typical Enhancers do not retain the same open epigenetic signature in EpiSCs. Super Enhancers and their associated gene promoters acquire CpG methylation in EpiSCs. With regards to Typical Enhancers, half appear highly CpG methylated in ESCs, but the unmethylated portion do acquire CpG methylation in EpiSCs similarly to Super Enhancers. In both instances this implies a repressive chromatin state establishes at these regulatory regions as pluripotency matures.

6.2 The subdivision: the two classes of Super Enhancelet in ESCs

6.2.1 Some ESC Super Enhancer constituent enhancer units maintain an open chromatin state during pluripotency maturation

Although ESC Super Enhancers overall gain CpG methylation in EpiSCs, some constituent enhancer units (affectionately dubbed Super Enhancelets) retain an unmethylated state both *in vitro* (figure 4.1) and *in vivo* (figure 4.2). As EpiSCs harbour dramatically higher global levels of CpG methylation,(129,130) epigenetic mechanisms must actively preserve the unmethylated state of PU. I termed these Super Enhancelets Persistently Unmethylated (PU) and those that acquire CpG methylation Differentially Methylated (DM). Curiously, PU Super Enhancelets not only show minimal or relative hypomethylation compared to DM in pluripotent cell types, but also in extra-embryonic, and in more differentiated cell types (figure 4.3). Furthermore, PU compared to DM Super Enhancelets show a clear increase in normalised CpG fraction. The genome-wide depletion of CpG dinucleotides (41–43,48,49) results from the increased mutability of 5-mC,(50–52) which readily deaminate to yield thymine nucleotides.(53) If PU compared to DM CpGs more often exist unmethylated, this could explain their difference in CpG density. One could pursue this premise, by downloading and processing, and analysing all publicly available mouse bisulphite-seq and MeDIP-seq data to see whether PU relative to DM appear hypomethylated in diverse cell types throughout and beyond development.

Surprisingly, PU, DM and PM Super Enhancer subregions recapitulate their patterns of CpG methylation during pluripotency prior to fertilisation (figure 4.3D). Correlation testing revealed that oocyte CpG methylation profiles most closely resemble that of E3.5 ICM. This observation adds to a growing body of evidence(266,275,317–319) that oocyte CpG methylation patterns persist in the blastocyst. This implies that these regions either resist the global epigenetic reprogramming events that occur during early embryogenesis, or that they retain other epigenetic mechanisms that allows pluripotency to re-establish these patterns of CpG methylation. One could explore these hypotheses by conducting whole-genome bisulphite-seq and ChIP-seq for key chromatin marks (e.g. H3K27ac, H3K27me3, H3K4me1, H3K4me3) of the oocyte, zygote, and at multiple time points until gastrulation.

Super Enhancelets display distinct patterns of transcription factor binding throughout pluripotency (figure 4.6). In ESCs PU compared to DM bind higher densities of OCT4, OTX2 and P300, as well as maintaining more accessible chromatin, as assayed by FAIRE-seq. This suggests that, in ESCs, PU compared to DM form a denser – and perhaps more stable – hub of transcriptional regulation proteins. Whether cause or consequence – or both – of the increased accessibility of PU is unclear. In EpiLCs, the transcription factor binding density and chromatin accessibility of DM reduces to that of the PM. In contrast, PU maintain relatively high transcription factor binding and chromatin accessibility. Together with their actively maintained unmethylated state in EpiSCs, the protein recruitment profiles of PU in EpiLCs suggest they continue to act as enhancers under the primed pluripotency transcriptional network. In contrast, the DM decommission, as evidenced by their acquisition of CpG methylation, reduced transcription factor binding, and increased nucleosome-occupancy (inferred from FAIRE-seq).

6.2.2 During pluripotency maturation, PU-containing ESC Super Enhancers maintain protein binding and transcription of their associated genes

As transcription is the functional output of enhancer activity, to truly determine whether DM Super Enhancelets decommission as pluripotency matures one must assay the expression of

their associated genes. Super Enhancers contain either only PU, PU and DM, or only DM Super Enhancelets (see figures 4.1A and 4.7A). We found that those Super Enhancers that contain one or more PU (i.e. Maintained Super Enhancers) maintain expression of their associated genes, whereas those that only contain DM (i.e. Silenced Super Enhancers) silence expression of their associated genes in epiblast both *in vitro* and *in vivo* (figure 4.7). Although compelling, this result does not evidence that these patterns of transcription occur as a direct result of the re-organisation of the transcriptional network from naïve to primed pluripotency. To verify this is the case, we examined the expression of Maintained and Silenced Super Enhancer associated genes during ESC-to-EpiSC conversion. We saw a gradual, step-wise reduction in expression of Silenced Super Enhancer associated genes during ESC-to-EpiSC conversion (figure 4.8), confirming that DM decommission with the establishment of the primed pluripotency transcriptional network.

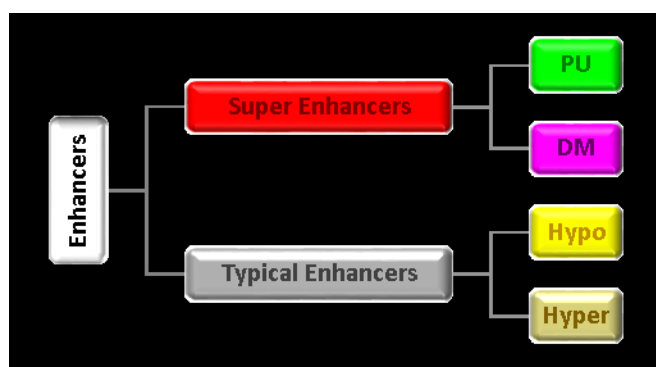


Figure 6.2 The four classes of ESC enhancer.

This story began with two classes of enhancer: Super Enhancers and Typical Enhancers. These classes were defined purely by MED1 ChIP-seq density as a proxy for Mediator complex occupation.(179) In chapters 4 and 6.1 I described how Typical Enhancers divide into two subclasses based on CpG methylation and transcription factor binding, called here ‘Hypo’ and ‘Hyper’ in reference to their relative CpG methylation profiles. This echoes the recent findings of Moorthy *et al.* who show that some Typical Enhancers exert a similar effect on transcription as Super Enhancers.(320) In chapter 4 and here, I show that the constituent enhancers within Super Enhancers also comprise two subclasses: PU and DM. Based on their

selective acquisition of CpG methylation, reduced transcription factor binding, increased nucleosome occupancy, and the transcriptional downregulation of genes associated with Super Enhancers containing them, only DM Super Enhancers decommission in EpiSCs and epiblast. PU Super Enhancers continue to maintain an active epigenetic signature, recruit relevant transcription factors, and drive the expression of their associated genes. This leaves us with, three – potentially four – classes of enhancers in ESCs: the PU and DM of Super Enhancers, and the Hypo and Hyper of Typical Enhancers (figure 6.2). In short, while the categorisation of any transcriptional phenomenon relies on arbitrary cut-offs to some extent, without a biology-led approach that integrates multiple types of information, one may oversimplify and obscure the underlying mechanisms.

6.3 The fulcrum and the lever: H3K4me3 balances TET/DNMT at PU and DM

Although traditionally thought of as a mark of promoters, I found high density H3K4me3 at PU in ESCs (figure 4.5). Several studies report the mutual exclusion of H3K4me3 and DNMTs.(284–287) Others report an associative relationship between H3K4me3 and the TET proteins.(278,321,322) (see schematic summary in figure 6.3A). Thus the high density of H3K4me3 at PU may inhibit the binding of DNMTs and promote local TET activity, in agreement with the higher TET1 occupancy relative to DNMT3A/B observed at PU (figure 4.4C). TET1 preferentially binds to CpG-rich DNA,(278,323) thus the increased CpG density of PU compared to DM (figure 4.1D) may contribute to the differential TET1 recruitment. Although PU and DM display a similar paucity of 5-mC in ESCs (serum+LIF), DM show a marked enrichment for 5-hmC (figure 4.4A). This observation, coupled with the increased ratio of TET1 to DNMT3A/B occupancy, led to the construction of a model (figure 6.3B) where:

1. DNA methylation events occur less frequently at PU compared to DM, as DNMTs relative to TETs have less contact time with DNA;

2. When DNA methylation events do occur at the PU, the high density of TET induces rapid demethylation through the mCpG oxidation cycle.

Under this model, DM would appear more densely marked for 5-hmC, due to the lower TET to DNMT binding ratio. The relative stability of 5-hmC could have mechanistic consequences for DM activity, for example, in the recruitment of transcription factors. In chapter 5 we observed that CTCF replicates the TET protein binding pattern at PU, DM and PM (i.e. similar levels of binding at DM and PM, compared to high levels at PU) (figure 5.3B), and that its DNA binding motif enriches at PU relative to DM (figure 5.2B). Study of CTCF at the imprinted H19/Igf2 locus shows methylation of the CpG within its binding motif disrupts its interaction with DNA.(100,101) 5-hmC may act similarly, although two publications report a positive association between CTCF and 5-hmC.(324,325) However, the exact relationship between CTCF and 5-hmC remains unclear. For example, the current published literature does not evidence a direct interaction between the two. Thus, the enrichment of 5-hmC at CTCF bound sites may reflect an active demethylation process within a heterogeneous cell population, and in fact 5-hmC does inhibit CTCF interacting with DNA. Alternatively, the perceived association between CpG methylation status and CTCF binding profile may result from the co-localisation of TET1 and CTCF in ESCs.(324) Researchers have documented enzymatically active TET1 interacting with CTCF in differentiating adipocytes.(325) Thus, TET1 and CTCF may recruit one another to PU and co-operate to an unknown end. As another possibility, 5-hmC may yet regulate protein binding at DM, as it interacts with numerous transcription factors in different cell types including ESCs.(94–99) Thus, 5-hmC may recruit or maintain the presence of unknown proteins at the DM that contribute to the differential behaviour of DM and PU (see figure 6.3A for schematic). As to the potential function of these molecules, combined TET1, TET2, and TET3 knock out ESCs, completely devoid of 5-hmC, maintain normal morphology, express OCT4 and NANOG, and form embryoid bodies, implying normal self-renewal and pluripotency.(326) However, they cannot contribute to chimeric embryonic and prove embryonically lethal. Thus the 5-hmC may not maintain DM, but instead choreograph their

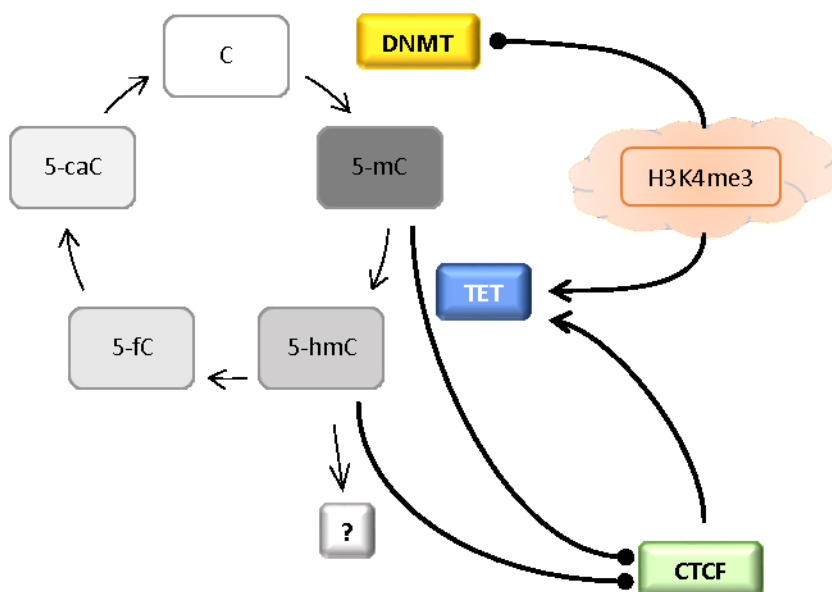
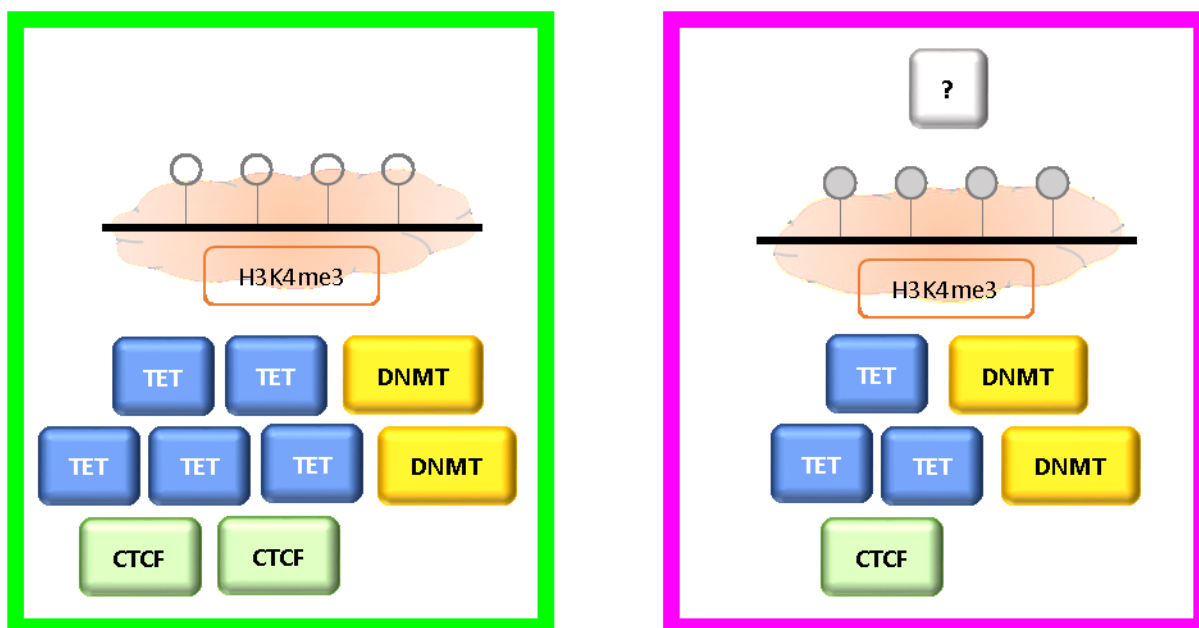
A**B**

Figure 6.3 The H3K4me3-TET-DNMT axis.

A. H3K4me3 promotes TET and inhibits DNMT recruitment to Super Enhancers thus modulating the 5-mC oxidation cycle. 5-mC and 5-hmC inhibit the binding of CTCF, which in turn regulates TET content. **B.** The consequences of differing levels of H3K4me3 at ESC Super Enhancers. The lollipops indicate CpGs (white – unmethylated, grey – hydroxymethylated).

decommissioning, enabling the smooth transition to primed pluripotency necessary for onward development.

Collectively, I propose that H3K4me3 acts as a fulcrum to the TET-DNMT lever at ESC Super Enhancelets. H3K4me3 promotes TET and inhibits DNMT localisation to these regions (figure 6.3A). At PU the high H3K4me3 content shifts the lever towards TET, leading to fewer CpG methylation events and rapid removal of 5-mC. At DM the relatively lower H3K4me3 content results in a more balanced ratio of TET and DNMT, which favours 5-hmC stability. The increased 5-hmC content of DM may in turn inhibit CTCF binding, perhaps reducing the stability of enhancer-promoter interactions, and/or recruit other transcription factors (see figure 6.3B).

To test the validity of this model, one could apply unsupervised machine learning techniques (e.g. clustering and PCA) to ESC ChIP-seq data to identify proteins that exhibit similar distribution patterns to H3K4me3, TET, and DNMT at PU and DM, and more widely at distal genomic regions. One could then use classical genetic depletion and over-expression models to examine the role of these proteins on protein-protein (co-immunoprecipitation), protein-DNA interactions (ChIP-seq and ChIP-qPCR), and epigenetic signatures (ChIP-seq for H3K4me3, sequencing techniques for base resolution of 5-hmC such as TAB-seq(327) and oxBS-seq(328)). A more elegant experimental model could employ the CRISPR-Cas9 system(81–88) to target TET1 and DNMT3A delivery(329) to candidate enhancers, thus perturbing the TET-DNMT balance locally rather than genome-wide. Such a model would allow focused study of this specific aspect of TET-DNMT-H3K4me3 at ESC-specific Super Enhancelets without interference from their activities in other contexts.

6.4 The linchpin: DM activity relies on ESRRB protein and co-regulators

Auto-regulatory, feed-forward loops(19) control ESC transcriptional network, which enable the buffering of small changes in concentration of key regulators.(330) However, dramatic shifts

in the balance of pluripotency transcription factors spark the destabilisation of the network, which *in vivo* induce differentiation. For example, we confirmed here that the constitutive deletion of *Esrrb* in ESCs compromises the plating efficiency of mutant cells, with self-renewal partly sustained through compensation mechanisms (figure 5.4B). Given such a delicate and complex system, it follows that of 100 key ESC transcription factors, depletion of 89 appears to deregulate the expression of Super Enhancer associated genes (figure 5.1D, table 5.2). Albeit, confirmation of these changes in gene expression would require RT-qPCR validation. In contrast, few of these transcription factors appear to selectively regulate Maintained (PU-containing) or Silenced (exclusively DM-containing) Super Enhancers (figure 5.1D-F, table 5.2). Of those, only ESRRB, its co-activator NCOA3, and CTCF also appear selectively enriched in terms of DNA binding motif (not NCOA3) (figure 5.2B) and protein binding density (figure 5.3). We chose to explore a potential role for ESRRB and NCOA3 in directing DM activity.

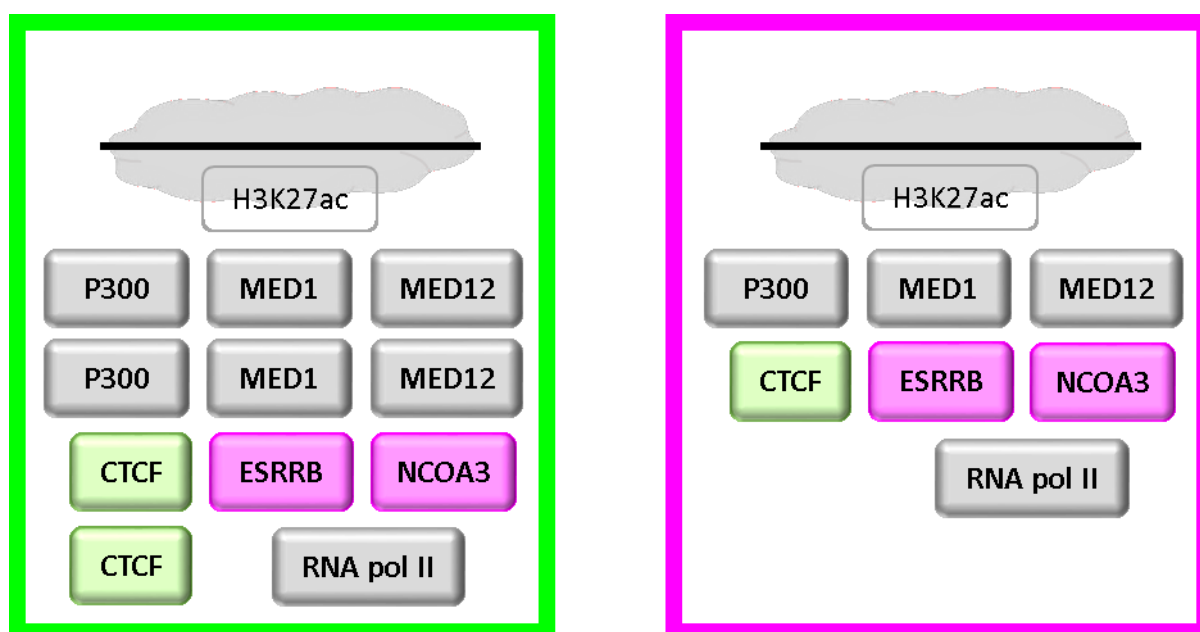


Figure 6.4 The differing balance between *ESRRB*-*NCOA3* and core transcriptional regulators at PU and DM.

Microarray meta-analyses of *Esrrb* and *Nanog* depleted ESCs, validates that *Esrrb* loss selectively impacts Silenced Super Enhancer associated genes (figure 5.5). This implies that DM compared to PU suffer the loss of *Esrrb* more keenly, which supports the hypothesis that ESRRB acts as a linchpin of the DM enhancer machinery. In contrast *Nanog* depletion shows

no selective effect on expression. These observations were validated in constitutive knock out ESC models (figure 5.6). *Esrrb* and *Ncoa3* exhibit a clear downregulation of Silenced Super Enhancer associated genes, whereas Maintained Super Enhancer associated genes retain a median $\log_2(\text{fold-change}) \sim 0$. Rescuing ESCs^{*Esrrb*^{-/-}} with wild-type *Esrrb* restores the expression of Silenced Super Enhancer associated genes (figure 5.8). The same rescue performed with *Esrrb*ΔAF-2 point-mutant form further reduces the expression of both classes of Super Enhancer associated gene. Given that ESRRB interacts with NCOA3 via the AF-2 domain, this evidences that the expression of Silenced Super Enhancer associated genes depends on NCOA3 working in conjunction with ESRRB. Figure 6.4 summarises my model of ESRRB-NCOA3 and core transcriptional regulator binding at PU and DM. I propose two possible mechanistic explanations for the role of ESRRB-NCOA3 at DM:

1. The differing balance between ESRRB-NCOA3 and core transcriptional regulators (e.g. the Mediator complex) at PU and DM, leaves the DM in a fragile state easily unravelled as the primed pluripotency transcriptional network emerges.
2. ESRRB forms part of a DM-specific regulatory network.

Although at present we lack the data to confirm either of these hypotheses, there are clear lines of enquiry to pursue.

6.4.1 Does the stability of DM rely on the presence of ESRRB-NCOA3?

During differentiation ESRRB and NCOA3 expression gradually declines.(301) I propose that the loss of ESRRB and NCOA3 selectively destabilises DM in the same way that the loss of a leg more greatly impacts a three-legged stool compared to a four-legged stool. The massive transcription factor binding, highly accessible chromatin, and CTCF-stabilised promoter interactions, enables the PU to continue its activity as the primed pluripotency network establishes, whereas the DM collapse. To pursue this hypothesis, one would need to destabilise the PU and stabilise the DM, then examine the effect on transcription factor recruitment, their epigenetic signature, expression of Maintained and Silenced Super

Enhancer associated genes, and the establishment of the primed pluripotency transcriptional network.

Earlier this year Kubo *et al.*, reported generating an ESC line with almost total CTCF depletion,(331) which would provide one method of destabilising the PU. They found that in the absence of CTCF, chromatin interactions appear weaker, but persist, along with a normal ESC transcriptome and epigenome. PU within these CTCF depleted cells may exist in a more fragile state. shRNA depletion or small molecule antagonism of ESRRB and/or NCOA3 in these cells may impact the expression of Maintained as well as Silenced Super Enhancer associated gene sets. Conversely, CRISPR-Cas9 targeting of CTCF to DM may promote their stabilisation, leading to a lesser effect on Silenced Super Enhancer associated genes upon ESRRB depletion. Another possible method arises from the recent findings of Oldfield *et al.*, who reported that the transcriptional co-regulator NF-Y promotes chromatin accessibility and recruitment of OSN to ESC-specific enhancers including the OCT4 Super Enhancer.(332) Targeted delivery of NF-Y to DM may promote local chromatin accessibility and transcription factor recruitment, stabilising the enhancer. In contrast, targeted delivery of a non-functional mutant of NF-Y to PU could reduce chromatin accessibility and OSN recruitment. These experiments could help clarify whether the increased ratio of ESRRB-NCOA3 to general pluripotency factors and chromatin regulators at DM compared to PU, results in a larger role for ESRRB-NCOA3 in the activity of DM. Thus, as ESRRB-NCOA3 depletes during pluripotency maturation, the DM protein network collapses leading to their decommissioning.

6.4.2 Does ESRRB form part of a DM-specific regulatory network?

We know ESRRB interacts with many proteins in ESCs independently of OCT4 (see table 7.1 for summary),(249) and that ESRRB interacts with RNA polymerase II via NCOA3.(301) Therefore ESRRB could feasibly regulate a DM-specific network of transcription factors, which dissolves in its absence. The chromatin modifying NuRD complex presents one such candidate as many of its members interact with ESRRB in an OCT-independent manner. Earlier this year Ee *et al.*, identified MBD3C, an isoform of the NuRD complex member MBD3,

which interacts with the TRX/MLL complex member WDR5 in ESCs.(333) ESCs abundantly express MBD3C and its expression downregulates during embryoid body differentiation. No other isoform of MBD3 interacts with WDR5. ESRRB interacts with both MBD3 and WDR5 independently of OCT4 (table 7.1).(249) The documented physical interactions of these three molecules buoys the possibility of a regulatory relationship. Moreover, MBD3 proves critical to the reprogramming of EpiSCs to iPSCs and therefore the establishment of the naïve pluripotency transcriptional network.(334)

To identify potential members of this hypothetical network, one could perform ChIP-Mass Spectrometry (MS) for MED1 in the ESC^{Esrrb-/-} and ESC^{Esrrb+/-} to identify proteins that interact with MED1 in an ESRRB-dependent manner. Using MED1 as the bait would enrich for Super Enhancer constituent enhancers, which would allow us to focus on this facet of ESRRB. As NCOA3 appears vital to the role of ESRRB at DM one could supplement this experiment was ChIP-MS for MED1 in ESC^{Esrrb-/-|Esrrb} and ESC^{Esrrb-/-|EsrrbΔAF-2}. One could validate putative interaction partners using co-immunoprecipitation (Co-IP) in these cells. A similar, but more focused experiment could use CRISPR-Cas9 to target ESRRBΔAF-2 to a candidate DM, followed by ChIP-qPCR for the putative interaction partner.

Table 6.1 OCT4-independent ESRRB interaction partners.

Proteins found to interact with ESRRB in an OCT4-independent manner according to ChIP-Mass Spectrometry (MS) analysis conducted by Dr Raymond Poot's group.(249) Class/complex – the class of proteins or complex, Protein – the proteins for within that class/complex that were found to interact. TF – Transcription Factor.

Class/ complex	Protein
SWI/SNF	ARID1A, PBRM1
Trrap/p400	YEATS4
NuRD	MTA2, GATAD2B, MTA3, MBD3, HDAC1, HDAC2, RBBP4
Mediator	MED14, MED12, MED23, MED17, MED1, MED15, MED13, MED13L, MED4, MED26, MED8, CNCC, MEC29, CDK8, MED19
RNApol2	POLR2B, POLR2C, POLR2G
TFIID	TAF9
TRX/MLL	WDR5, MLL3, ASHL2, RBBP5
TF	FKBP15, ESRRA, UBP1, CDC2A, ZBTB9, REQUIEM, TCFCP2, MYST1

6.5 The beacon: enhancers in murine pluripotent cell types

Four years after their publication, researchers remain divided over the biological novelty of Super Enhancers and the related concepts of Seed(283) and Stretch(335) Enhancers. Many researchers discussing Super Enhancers informally on social networking websites(336–339) and magazine commentaries(340) appear dismissive. Pott and Leib published the first formal rebuttal to Super Enhancers in 2014.(341) They criticise their definition on the ground that “there does not seem to be a biological rationale or motivation underlying the specific cut-offs used in the definition of super-enhancers”. Regarding their novelty they state that “enhancers that comprise super-enhancers exhibit previously known characteristics of mammalian enhancers ... many of these characteristics were established in regions discovered in the 1970s through 1990s before the advent of genome-wide studies.” However, they note that the methods of identifying Super Enhancers will enable researchers to quickly identify cell-specific regulatory regions in their models of interest. They conclude by warning that to use Super Enhancers to study other biological phenomena when their biological novelty and definition remains unclear, risks falling into a circular logic trap. This year Hamdan and Johnsen describe the definitions of Super and Typical Enhancers as “still poorly defined and fairly arbitrary”, and echo Pott and Leib’s counsel that “SE[s] may simply be ‘low hanging fruit’ of enhancers that can prejudice our investigations and prevent us from identifying other important regulatory elements”.(342) In agreement, Moorthy et al., state that “the super-enhancer prediction model ignores a large number of functional regulatory regions in ES cells and overemphasizes the importance of clustered enhancers compared to isolated enhancers.”(320) Nonetheless, the definition and biological validity proved sufficient to issue a patent on Prof Young’s method and future applications of Super Enhancers.(343)

The number of related publications increases each year (figure 6.5). At the time of writing, researchers have this year published 85 PubMed and 81 bioRxiv papers related to Super

Enhancers. These publications both support and detract from the original Super Enhancer concept. Moorthy *et al.*, define their own set of ESC Super Enhancers by integrating EP300, H3K4me1, MED12 and NIPBL ChIP-seq data with previously defined multiple transcription factor bound loci (MTLs).(320) To pair their Super Enhancers with genes, they performed RNA-seq on ESCs carrying heterozygous deletions of candidate Super Enhancer constituent enhancers. Although much more likely to accurately identify active enhancer-promoter, this approach prevents them from analysing global expression patterns. They found that constituent enhancers vary greater in their contribution to regulation of their target genes, some constituent enhancers act in a partially redundant manner, some constituent enhancers exert no effect on expression of the gene they target (consistent with the idea of a poised enhancer-promoter interaction(189)). A similarly meticulous dissection of a Super Enhancer driving *Wap* expression in mammary reports complementary findings.(344) Xie *et al.*, went a step further by dissecting Super Enhancers using their Mosaic-seq technique, which uses the CRISPR-Cas9 system to deplete a one constituent enhancer per cell, followed by single-cell RNA-seq.(345) They found that of the four constituent enhancers within the β -Globin locus control region (LCR) (which shares similar properties to Super Enhancers), deletion of any reduced transcription of HBG2, but loss of constituent enhancer HS2 invoked a substantially greater loss. Notably, Moorthy *et al.*, did not find that deletion of Super Enhancer compared to Typical Enhancer constituents affected gene expression more strongly. Furthermore, when they analysed the 231 original ESC Super Enhancers, they found that 28 of them did not contain an MTL or any of their predicted enhancer regions, possibly meaning they do not represent an enhancer of any type. Using BRD4 occupancy, Hamdan and Johnsen define 324 Super and 24,427 Typical Enhancers in the MCF7 breast cancer cell line.(342) To narrow down the list of Typical Enhancers, they intersected them with ER α occupation sites. BRD4 depletion triggered the downregulation of twice as many Typical compared to Super Enhancer associated genes, leading Hamdan and Johnsen to question whether the Typical Enhancers ought to be overlooked. Together this work evidences the subjective and biologically uninformed nature of the original Super Enhancer definition.

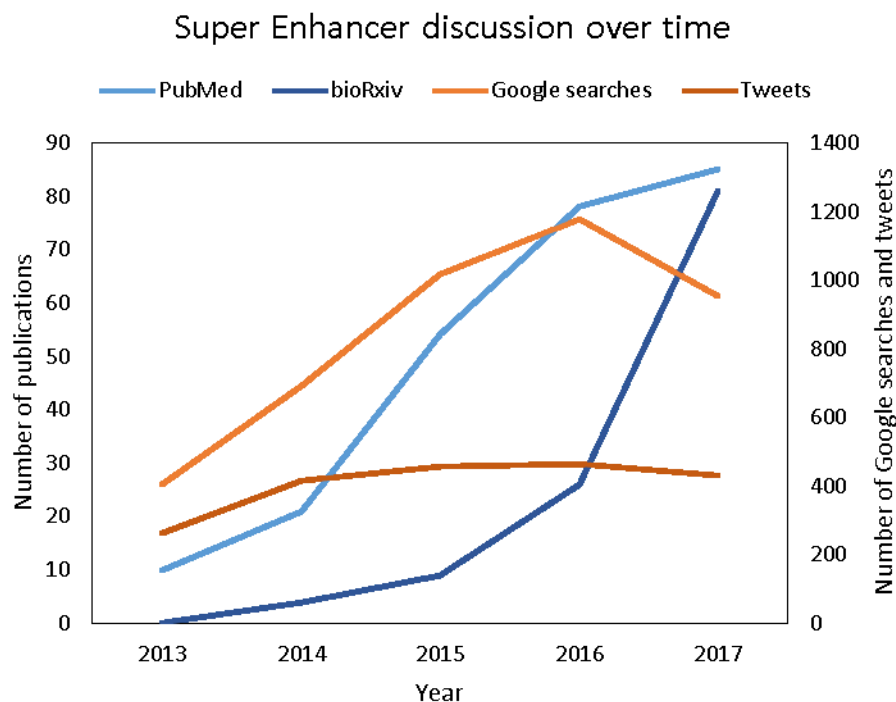


Figure 6.5 The formal (PubMed and bioRxiv) and informal (Google searches and tweets) discussion of Super Enhancers.

Although heavily criticised, the concept of Super Enhancers continues to generate interest.

In conclusion, the enhancers within mouse ESCs behave highly heterogeneously. They differ in their capacity to drive gene expression in ESCs and specificity to the different stages of pluripotency. As we have seen here, any attempt to categorise genomic regulatory elements relies on arbitrary cut-offs – informed by our preconceptions as much as biology – that can over-simplify and obscure the underlying genetic and epigenetic complexity of these regions. The concept of Super Enhancers and Typical Enhancers presents a false dichotomy, resulting from methodological limitations. Some Typical Enhancers drive transcription equally as well as Super Enhancer constituents. Furthermore, the constituent enhancers within ESC Super Enhancers are not identical; they recruit transcription factors at different densities, bear distinct epigenetic marks, and befall different fates upon pluripotency maturation. Ignorance of the methodological limitations in the definition of Super Enhancers and Typical Enhancers may collectively hinder the field from furthering our understanding of transcriptional regulation of

gene expression. However, based on the data generated for this thesis, in a given cell type MED1 or H3K27ac binding density alone clearly can accurately define clusters of enhancers that actively drive transcription. The simplicity of this method empowers researchers who lack the resources to generate vast swathes of sequencing data, use uncommon model systems, or work *in vivo*, to identify enhancers from a single ChIP-seq experiment. In 2008 the plummeting cost of sequencing technology sparked a renaissance in enhancer research (figure 1.3). Perhaps our new-found capacity to simply and accurately define enhancer clusters in any given eukaryotic cell type will do so too.

BIBLIOGRAPHY

1. Adams D. Mostly harmless. Ballantine Books; 1993. 277 p.
2. Bernstein R. Depression afflicts almost half of STEM graduate students at UC Berkeley. *Science* (80-) [Internet]. 2015 May 13 [cited 2017 Nov 7]; Available from: http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2015_05_13/caredit.a1500125
3. Mind. How common are mental health problems? | Mind, the mental health charity - help for mental health problems [Internet]. [cited 2017 Nov 7]. Available from: <https://www.mind.org.uk/information-support/types-of-mental-health-problems/statistics-and-facts-about-mental-health/how-common-are-mental-health-problems/>
4. Random House. Definition: Totipotent. Dictionary.com Unabridged [Internet]. 2017; Available from: <http://www.dictionary.com/browse/totipotent>
5. Kojima Y, Kaufman-Francis K, Studdert JB, Steiner KA, Power MD, Loebel DAF, et al. The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell* [Internet]. 2014 Jan 2 [cited 2013 Nov 10];14(1):107–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24139757>
6. Nichols J, Smith A. Naive and primed pluripotent states. *Cell Stem Cell* [Internet]. 2009 Jun 5 [cited 2014 Jul 9];4(6):487–92. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19497275>
7. Weinberger L, Ayyash M, Novershtern N, Hanna JH. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol* [Internet]. *Nature Research*; 2016 Feb 10 [cited 2017 Oct 27];17(3):155–69. Available from: <http://www.nature.com/doifinder/10.1038/nrm.2015.28>
8. Niwa H, Burdon T, Chambers I, Smith A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* [Internet]. 1998 Jul 1 [cited 2015 Jun 5];12(13):2048–60. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=316954&tool=pmcentrez&rendertype=abstract>
9. Ying QL, Nichols J, Chambers I, Smith A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* [Internet]. 2003 Oct 31 [cited 2015 Apr 21];115(3):281–92. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14636556>
10. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* [Internet]. 2007 Jul 12 [cited 2013 Sep 17];448(7150):196–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17597760>
11. Brons IGM, Smithers LE, Trotter MWB, Rugg-Gunn P, Sun B, Chuva de Sousa Lopes SM, et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* [Internet]. 2007 Jul 12 [cited 2013 Sep 23];448(7150):191–5. Available from: <http://dx.doi.org/10.1038/nature05950>
12. Torres-Padilla M-E, Chambers I. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *Development* [Internet]. Oxford University Press for The Company of Biologists Limited; 2014 Jun 1 [cited 2017 Oct 26];141(11):2173–81.

- Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24866112>
13. Toyooka Y, Shimosato D, Murakami K, Takahashi K, Niwa H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* [Internet]. The Company of Biologists Ltd; 2008 Mar 1 [cited 2017 Oct 27];135(5):909–18. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1821852>
 14. Ying Q-L, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, et al. The ground state of embryonic stem cell self-renewal. *Nature* [Internet]. Nature Publishing Group; 2008 May 22 [cited 2015 Sep 15];453(7194):519–23. Available from: <http://dx.doi.org/10.1038/nature06968>
 15. Wray J, Kalkan T, Smith AG. The ground state of pluripotency. *Biochem Soc Trans* [Internet]. 2010 Aug [cited 2015 Apr 27];38(4):1027–32. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20658998>
 16. Young RA. Control of the embryonic stem cell state. *Cell* [Internet]. Elsevier; 2011 Mar 18 [cited 2013 Sep 17];144(6):940–54. Available from: [http://www.cell.com/fulltext/S0092-8674\(11\)00071-7](http://www.cell.com/fulltext/S0092-8674(11)00071-7)
 17. Kim J, Chu J, Shen X, Wang J, Orkin SH. An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell* [Internet]. Elsevier; 2008 Mar 21 [cited 2016 Jun 29];132(6):1049–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18358816>
 18. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* [Internet]. 2008 Jun 13 [cited 2013 Dec 16];133(6):1106–17. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18555785>
 19. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* [Internet]. Cell Press; 2005 Sep 23 [cited 2017 Oct 19];122(6):947–56. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867405008251?via%3Dihub>
 20. Loh Y-H, Wu Q, Chew J-L, Vega VB, Zhang W, Chen X, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* [Internet]. 2006 Apr 5 [cited 2015 Jun 18];38(4):431–40. Available from: <http://dx.doi.org/10.1038/ng1760>
 21. Schöler HR, Dressler GR, Balling R, Rohdewohld H, Gruss P. Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J* [Internet]. 1990 Jul [cited 2017 Oct 27];9(7):2185–95. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2357966>
 22. Okamoto K, Okazawa H, Okuda A, Sakai M, Muramatsu M, Hamada H. A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells. *Cell* [Internet]. Cell Press; 1990 Feb 9 [cited 2017 Oct 27];60(3):461–72. Available from: <http://www.sciencedirect.com/science/article/pii/0092867490905978?via%3Dihub>
 23. Rosner MH, Vigano MA, Ozato K, Timmons PM, Poirie F, Rigby PWJ, et al. A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. *Nature* [Internet]. Nature Publishing Group; 1990 Jun 21 [cited 2017 Oct 27];345(6277):686–92. Available from: <http://www.nature.com/doifinder/10.1038/345686a0>
 24. Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, et al. Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU

- Transcription Factor Oct4. *Cell* [Internet]. Cell Press; 1998 Oct 30 [cited 2017 Jul 11];95(3):379–91. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867400817699>
25. Niwa H, Smith AG, Miyazaki J. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* [Internet]. Nature Publishing Group; 2000 Apr 1 [cited 2017 Jul 11];24(4):372–6. Available from: <http://www.nature.com/doi/10.1038/74199>
 26. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* [Internet]. Cold Spring Harbor Laboratory Press; 2003 Jan 1 [cited 2017 Jul 11];17(1):126–40. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12514105>
 27. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol* [Internet]. Nature Publishing Group; 2007 Jun 21 [cited 2017 Oct 27];9(6):625–35. Available from: <http://www.nature.com/doi/10.1038/ncb1589>
 28. Kopp JL, Ormsbee BD, Desler M, Rizzino A. Small Increases in the Level of Sox2 Trigger the Differentiation of Mouse Embryonic Stem Cells. *Stem Cells* [Internet]. John Wiley & Sons, Ltd.; 2008 Apr 1 [cited 2017 Oct 27];26(4):903–11. Available from: <http://doi.wiley.com/10.1634/stemcells.2007-0951>
 29. Chew J-L, Loh Y-H, Zhang W, Chen X, Tam W-L, Yeap L-S, et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol* [Internet]. American Society for Microbiology; 2005 Jul 1 [cited 2017 Jul 11];25(14):6031–46. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15988017>
 30. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, et al. Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. *Cell* [Internet]. Cell Press; 2003 May 30 [cited 2017 Jul 11];113(5):643–55. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867403003921>
 31. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* [Internet]. 2003 May 30 [cited 2017 Jul 11];113(5):631–42. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12787504>
 32. Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, et al. Nanog Is the Gateway to the Pluripotent Ground State. *Cell* [Internet]. Cell Press; 2009 Aug 21 [cited 2017 Oct 27];138(4):722–37. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867409009696?via%3Dihub>
 33. Guo G, Yang J, Nichols J, Hall JS, Eyres I, Mansfield W, et al. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* [Internet]. 2009 Apr [cited 2013 Sep 24];136(7):1063–9. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2685927&tool=pmcentrez&rendertype=abstract>
 34. Buecker C, Srinivasan R, Wu Z, Calo E, Acampora D, Faial T, et al. Reorganization of Enhancer Patterns in Transition from Naive to Primed Pluripotency. *Cell Stem Cell* [Internet]. 2014 Jun 5 [cited 2014 Jun 5];14(6):838–53. Available from: <http://www.sciencedirect.com/science/article/pii/S1934590914001404>
 35. Dunn S-J, Martello G, Yordanov B, Emmott S, Smith AG. Defining an essential transcription factor program for naïve pluripotency. *Science* [Internet]. 2014 Jun 6 [cited

- 2015 Oct 2];344(6188):1156–60. Available from: <http://www.sciencemag.org/content/344/6188/1156.abstract>
36. Qiu D, Ye S, Ruiz B, Zhou X, Liu D, Zhang Q, et al. Klf2 and Tfcp2l1, Two Wnt/ β -Catenin Targets, Act Synergistically to Induce and Maintain Naive Pluripotency. *Stem Cell Reports* [Internet]. Elsevier; 2015 Sep 8 [cited 2017 Oct 2];5(3):314–22. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S2213671115002209>
 37. Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, et al. Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. *Cell Stem Cell* [Internet]. 2010 Mar 5 [cited 2015 Mar 3];6(3):215–26. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1934590910000044>
 38. Gillich A, Bao S, Grabole N, Hayashi K, Trotter MWB, Pasque V, et al. Epiblast Stem Cell-Based System Reveals Reprogramming Synergy of Germline Factors. *Cell Stem Cell* [Internet]. 2012 Apr [cited 2017 Oct 27];10(4):425–39. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1934590912000598>
 39. Hanna J, Markoulaki S, Mitalipova M, Cheng AW, Cassady JP, Staerk J, et al. Metastable Pluripotent States in NOD-Mouse-Derived ESCs. *Cell Stem Cell* [Internet]. 2009 Jun [cited 2017 Oct 27];4(6):513–24. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1934590909001696>
 40. Waddington CH. The Epigenotype. *Endeavor* [Internet]. Oxford University Press; 1942 Feb 1 [cited 2017 Oct 29]; Available from: <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyr184>
 41. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. *Science* (80-) [Internet]. 2001 [cited 2017 Jul 18];291(5507). Available from: <http://science.sciencemag.org/content/291/5507/1304>
 42. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* [Internet]. 2001 Feb 15 [cited 2017 Jul 18];409(6822):860–921. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11237011>
 43. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* [Internet]. Nature Publishing Group; 2002 Dec 5 [cited 2015 Jan 2];420(6915):520–62. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12466850>
 44. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science* [Internet]. 1974 May 24 [cited 2017 Oct 29];184(4139):868–71. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/4825889>
 45. Oudet P, Gross-Bellard M, Chambon P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* [Internet]. Cell Press; 1975 Apr 1 [cited 2017 Oct 29];4(4):281–300. Available from: <http://www.sciencedirect.com/science/article/pii/009286747590149X>
 46. Yokochi T, Robertson KD. Preferential methylation of unmethylated DNA by Mammalian de novo DNA methyltransferase Dnmt3a. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2002 Apr 5 [cited 2017 Oct 29];277(14):11735–45. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11821381>
 47. Ramsahoye BH, Biniszkiwicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 2000 May 9 [cited 2017 Oct 29];97(10):5237–42. Available from:

- <http://www.ncbi.nlm.nih.gov/pubmed/10805783>
48. Bird A. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* [Internet]. 1980 [cited 2013 Mar 10];(4):1499–504. Available from: <http://nar.oxfordjournals.org/content/8/7/1499.short>
 49. Nagy A, Rossant J, Nagy R, Abramow-Newerly W, Roder JC, Burge C, et al. Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proc Natl Acad Sci* [Internet]. BioMed Central; 1993 Sep 15 [cited 2017 Jun 27];90(18):8424–8. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.90.18.8424>
 50. Youssoufian H, Kazazian HH, Phillips DG, Aronis S, Tsiftis G, Brown VA, et al. Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. *Nature* [Internet]. 1986 Nov 27 [cited 2017 Jul 18];324(6095):380–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3097553>
 51. Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* [Internet]. 2008 Aug [cited 2017 Jul 18];29(8):1037–47. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18454449>
 52. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. *Hum Genet* [Internet]. 1988 Feb [cited 2017 Jul 18];78(2):151–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3338800>
 53. Shen JC, Rideout WM, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* [Internet]. 1994 Mar 25 [cited 2017 Jul 18];22(6):972–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8152929>
 54. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *J Mol Biol* [Internet]. 1987 Jul [cited 2014 Feb 11];196(2):261–82. Available from: <http://www.sciencedirect.com/science/article/pii/0022283687906899>
 55. Ponger L, Duret L, Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2001 Nov 1 [cited 2017 Jul 18];11(11):1854–60. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11691850>
 56. Takai D, Jones P a. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci* [Internet]. National Academy of Sciences; 2002 Mar 19 [cited 2017 Jul 18];99(6):3740–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11891299>
 57. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* [Internet]. 1992 Jun 12;69(6):915–26. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1606615>
 58. Bostick M, Kim JK, Estève P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* [Internet]. American Association for the Advancement of Science; 2007 Sep 21 [cited 2017 Oct 28];317(5845):1760–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17673620>
 59. Sharif J, Muto M, Takebayashi S, Suetake I, Iwamatsu A, Endo TA, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* [Internet]. Nature Publishing Group; 2007 Dec 6 [cited 2017 Oct 28];450(7171):908–12. Available from: <http://www.nature.com/doi/10.1038/nature06397>
 60. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion

- of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* (80-) [Internet]. NIH Public Access; 2009 May 15 [cited 2017 Jun 29];324(5929):930–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19372391>
61. Ito S, D'Alessio AC, Taranova O V., Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* [Internet]. Nature Research; 2010 Aug 26 [cited 2017 Oct 29];466(7310):1129–33. Available from: <http://www.nature.com/doi/10.1038/nature09303>
 62. He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* [Internet]. NIH Public Access; 2011 Sep 2 [cited 2017 Oct 29];333(6047):1303–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21817016>
 63. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* [Internet]. Howard Hughes Medical Institute; 2011 Sep 2 [cited 2017 Jun 29];333(6047):1300–3. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21778364>
 64. Weber AR, Krawczyk C, Robertson AB, Kuśnierczyk A, Vågbo CB, Schuermann D, et al. Biochemical reconstitution of TET1–TDG–BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat Commun* [Internet]. 2016 Mar 2 [cited 2017 Oct 29];7:10806. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26932196>
 65. Maiti A, Drohat AC. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine. *J Biol Chem* [Internet]. 2011 Oct 14 [cited 2017 Oct 29];286(41):35334–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21862836>
 66. Edwards JR, Yarychivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. *Epigenetics Chromatin* [Internet]. BioMed Central; 2017 [cited 2017 Oct 29];10:23. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28503201>
 67. McGhee J, Ginder G. Specific DNA methylation sites in the vicinity of the chicken β -globin genes. *Nature* [Internet]. Nature Publishing Group; 1979 Aug 2 [cited 2017 Jul 20];280(5721):419–20. Available from: <http://www.nature.com/doi/10.1038/280419a0>
 68. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* [Internet]. 1981 Dec [cited 2015 Jul 29];27(2 Pt 1):299–308. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6277502>
 69. McGhee J. A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell* [Internet]. 1981 Nov [cited 2017 Jul 20];27(1):45–55. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0092867481903597>
 70. Keshet I, Lieman-Hurwitz J, Cedar H. DNA methylation affects the formation of active chromatin. *Cell* [Internet]. 1986 Feb [cited 2017 Jul 20];44(4):535–43. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0092867486902631>
 71. Busslinger M, Hurst J, Flavell RA. DNA methylation and the regulation of globin gene expression. *Cell* [Internet]. 1983 Aug [cited 2017 Jul 20];34(1):197–206. Available from: <http://linkinghub.elsevier.com/retrieve/pii/0092867483901502>
 72. Kruczek I, Doerfler W. Expression of the chloramphenicol acetyltransferase gene in mammalian cells under the control of adenovirus type 12 promoters: effect of promoter methylation on gene expression. *Proc Natl Acad Sci U S A* [Internet]. 1983 Dec [cited 2017 Jul 20];80(24):7586–90. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6461111>

- <http://www.ncbi.nlm.nih.gov/pubmed/6324179>
73. Langner KD, Vardimon L, Renz D, Doerfler W. DNA methylation of three 5' C-C-G-G 3' sites in the promoter and 5' region inactivate the E2a gene of adenovirus type 2. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 1984 May [cited 2017 Jul 20];81(10):2950–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6328479>
 74. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011 Dec 22 [cited 2014 Jul 15];480(7378):490–5. Available from: <http://dx.doi.org/10.1038/nature10716>
 75. Schmidl C, Klug M, Boeld TJ, Andreessen R, Hoffmann P, Edinger M, et al. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2009 Jul 1 [cited 2017 Jul 20];19(7):1165–74. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19494038>
 76. Saluz HP, Jiricny J, Jost JP. Genomic sequencing reveals a positive correlation between the kinetics of strand-specific DNA demethylation of the overlapping estradiol/glucocorticoid-receptor binding sites and the rate of avian vitellogenin mRNA synthesis. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 1986 Oct [cited 2017 Jul 20];83(19):7167–71. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3463957>
 77. Wiench M, John S, Baek S, Johnson TA, Sung M-H, Escobar T, et al. DNA methylation status predicts cell type-specific enhancer activity. *EMBO J* [Internet]. EMBO Press; 2011 Aug 3 [cited 2014 Feb 11];30(15):3028–39. Available from: <http://emboj.embopress.org/content/30/15/3028.abstract>
 78. Bock C, Beerman I, Lien W-H, Smith ZD, Gu H, Boyle P, et al. DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Mol Cell* [Internet]. 2012 [cited 2017 Jul 20];47(4):633–47. Available from: <http://www.sciencedirect.com/science/article/pii/S1097276512005448>
 79. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol* [Internet]. 2013 Mar 12 [cited 2014 Apr 10];14(3):R21. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-3-r21>
 80. Magnusson M, Lu EX, Larsson P, Ulfhammer E, Bergh N, Carén H, et al. Dynamic Enhancer Methylation - A Previously Unrecognized Switch for Tissue-Type Plasminogen Activator Expression. El-Maarri O, editor. *PLoS One* [Internet]. Public Library of Science; 2015 Oct 28 [cited 2017 Jul 20];10(10):e0141805. Available from: <http://dx.plos.org/10.1371/journal.pone.0141805>
 81. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li G-W, et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* [Internet]. Cell Press; 2013 Dec 19 [cited 2017 Oct 19];155(7):1479–91. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867413015316>
 82. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* [Internet]. Cell Press; 2013 Feb 28 [cited 2017 Oct 19];152(5):1173–83. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867413002110>
 83. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, et al.

- Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* [Internet]. Nature Publishing Group; 2014 Dec 10 [cited 2017 Oct 19];517(7536):583–8. Available from: <http://www.nature.com/doi/10.1038/nature14136>
84. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* [Internet]. Nature Research; 2015 Apr 6 [cited 2017 Oct 19];33(5):510–7. Available from: <http://www.nature.com/doi/10.1038/nbt.3199>
 85. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* [Internet]. Cell Press; 2013 Jul 18 [cited 2017 Oct 19];154(2):442–51. Available from: <http://www.sciencedirect.com/science/article/pii/S009286741300826X>
 86. Horvath P, Barrangou R, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* (80-) [Internet]. American Association for the Advancement of Science; 2010 Jan 8 [cited 2017 Oct 19];327(5962):167–70. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20056882>
 87. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* (80-) [Internet]. American Association for the Advancement of Science; 2007 Mar 23 [cited 2017 Oct 19];315(5819):1709–12. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17379808>
 88. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, et al. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* (80-) [Internet]. American Association for the Advancement of Science; 2008 Aug 15 [cited 2017 Oct 19];321(5891):960–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18703739>
 89. Stepper P, Kungulovski G, Jurkowska RZ, Chandra T, Krueger F, Reinhardt R, et al. Efficient targeted DNA methylation with chimeric dCas9–Dnmt3a–Dnmt3L methyltransferase. *Nucleic Acids Res* [Internet]. Oxford University Press; 2017 Feb 28 [cited 2017 Oct 30];45(4):1703–13. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1112>
 90. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, et al. Editing DNA Methylation in the Mammalian Genome. *Cell* [Internet]. 2016 Sep [cited 2017 Oct 30];167(1):233–247.e17. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867416311539>
 91. Amabile A, Migliara A, Capasso P, Biffi M, Cittaro D, Naldini L, et al. Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. *Cell* [Internet]. 2016 Sep [cited 2017 Oct 30];167(1):219–232.e14. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867416312363>
 92. Vojta A, Dobrinić P, Tadić V, Bočkor L, Korać P, Julg B, et al. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res* [Internet]. Oxford University Press; 2016 Jul 8 [cited 2017 Oct 30];44(12):5615–28. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw159>
 93. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2017 Oct 16 [cited 2017 Oct 30];nrg.2017.80. Available from: <http://www.nature.com/doi/10.1038/nrg.2017.80>
 94. Wang D, Hashimoto H, Zhang X, Barwick BG, Lonial S, Boise LH, et al. MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic*

- Acids Res [Internet]. Oxford University Press; 2017 Mar 17 [cited 2017 Oct 18];45(5):2396–407. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1184>
95. Hashimoto H, Olanrewaju YO, Zheng Y, Wilson GG, Zhang X, Cheng X. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev* [Internet]. Cold Spring Harbor Laboratory Press; 2014 Oct 15 [cited 2017 Oct 18];28(20):2304–13. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25258363>
 96. Xiong J, Zhang Z, Chen J, Huang H, Xu Y, Ding X, et al. Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. *Mol Cell* [Internet]. 2016 Dec [cited 2017 Oct 18];64(5):913–25. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1097276516306360>
 97. Mellén M, Ayata P, Dewell S, Kriaucionis S, Heintz N. MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System. *Cell* [Internet]. 2012 Dec [cited 2017 Oct 18];151(7):1417–30. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867412014079>
 98. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PWTC, Bauer C, et al. Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. *Cell* [Internet]. 2013 Feb [cited 2017 Oct 18];152(5):1146–59. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867413001529>
 99. Iurlaro M, Ficiz G, Oxley D, Raiber E-A, Bachman M, Booth MJ, et al. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* [Internet]. BioMed Central; 2013 Dec 10 [cited 2017 Oct 18];14(10):R119. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-10-r119>
 100. Tilghman SM, Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* [Internet]. Nature Publishing Group; 2000 May 25 [cited 2017 Oct 19];405(6785):486–9. Available from: <http://www.nature.com/doi/10.1038/35013106>
 101. Felsenfeld G, Bell AC. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* [Internet]. Nature Publishing Group; 2000 May 25 [cited 2017 Oct 19];405(6785):482–5. Available from: <http://www.nature.com/doi/10.1038/35013100>
 102. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* (80-) [Internet]. American Association for the Advancement of Science; 2017 May 5 [cited 2017 Oct 15];356(6337):eaaj2239. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28473536>
 103. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, et al. DNA methylation presents distinct binding sites for human transcription factors. *Elife* [Internet]. eLife Sciences Publications Limited; 2013 Sep 3 [cited 2017 Oct 15];2:e00726. Available from: <http://elifesciences.org/lookup/doi/10.7554/eLife.00726>
 104. Liu Y, Olanrewaju YO, Zheng Y, Hashimoto H, Blumenthal RM, Zhang X, et al. Structural basis for Klf4 recognition of methylated DNA. *Nucleic Acids Res* [Internet]. Oxford University Press; 2014 Apr 1 [cited 2017 Oct 15];42(8):4859–67. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku134>
 105. Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet* [Internet]. Elsevier Current Trends; 2016 Jan 1 [cited 2017 Oct 30];32(1):42–56. Available from:

- <http://www.sciencedirect.com/science/article/pii/S0168952515001936>
106. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* [Internet]. Nature Publishing Group; 2011 Mar 15 [cited 2017 Oct 30];21(3):381–95. Available from: <http://www.nature.com/doifinder/10.1038/cr.2011.22>
 107. Kuo M-H, Allis CD. Roles of histone acetyltransferases and deacetylases in gene regulation. *BioEssays* [Internet]. 1998 Dec 7 [cited 2017 Oct 30];20(8):615–26. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9780836>
 108. Du Q, Luu P-L, Stirzaker C, Clark SJ. Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics* [Internet]. Future Medicine Ltd London, UK ; 2015 Sep 30 [cited 2017 Oct 30];7(6):1051–73. Available from: <http://www.futuremedicine.com/doi/10.2217/epi.15.39>
 109. Ng HH, Jeppesen P, Bird A. Active repression of methylated genes by the chromosomal protein MBD1. *Mol Cell Biol* [Internet]. American Society for Microbiology; 2000 Feb 15 [cited 2017 Oct 30];20(4):1394–406. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10648624>
 110. Wolffe AP, Wade PA, Gegonne A, Jones PL, Ballestar E, Aubry F. Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet* [Internet]. Nature Publishing Group; 1999 Sep 1 [cited 2017 Oct 30];23(1):62–6. Available from: <http://www.nature.com/doifinder/10.1038/12664>
 111. Boeke J, Ammerpohl O, Kegel S, Moehren U, Renkawitz R. The minimal repression domain of MBD2b overlaps with the methyl-CpG-binding domain and binds directly to Sin3A. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2000 Nov 10 [cited 2017 Oct 30];275(45):34963–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10950960>
 112. Agarwal N, Hardt T, Brero A, Nowak D, Rothbauer U, Becker A, et al. MeCP2 interacts with HP1 and modulates its heterochromatin association during myogenic differentiation. *Nucleic Acids Res* [Internet]. Oxford University Press; 2007 Aug 1 [cited 2017 Oct 30];35(16):5402–8. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm599>
 113. Nan X, Ng H-H, Johnson CA, Laherty CD, Turner BM, Eisenman RN, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* [Internet]. Nature Publishing Group; 1998 May 28 [cited 2017 Oct 30];393(6683):386–9. Available from: <http://www.nature.com/doifinder/10.1038/30764>
 114. Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T. The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2003 Feb 7 [cited 2017 Oct 30];278(6):4035–40. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12427740>
 115. Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res* [Internet]. Oxford University Press; 2012 Jun 1 [cited 2017 Oct 30];40(11):4841–9. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks155>
 116. Fraga MF, Ballestar E, Montoya G, Taysavang P, Wade PA, Esteller M. The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res* [Internet]. Oxford University Press; 2003 Mar 15 [cited 2017 Oct 30];31(6):1765–74. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg249>

117. Cramer JM, Scarsdale JN, Walavalkar NM, Buchwald WA, Ginder GD, Williams DC. Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2014 Jan 17 [cited 2017 Oct 30];289(3):1294–302. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24307175>
118. Yildirim O, Li R, Hung J-H, Chen PB, Dong X, Ee L-S, et al. Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. *Cell* [Internet]. 2011 [cited 2017 Jul 18];147(7):1498–510. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867411014498>
119. Saito M, Ishikawa F. The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2002 Sep 20 [cited 2017 Oct 30];277(38):35434–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12124384>
120. Baubec T, Ivánek R, Lienert F, Schübeler D. Methylation-dependent and -independent genomic targeting principles of the mbd protein family. *Cell* [Internet]. 2013 Apr 11 [cited 2017 Apr 27];153(2):480–92. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867413003334>
121. Cai Y, Geutjes E-J, de Lint K, Roepman P, Bruurs L, Yu L-R, et al. The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes. *Oncogene* [Internet]. Nature Publishing Group; 2014 Apr 24 [cited 2017 Oct 30];33(17):2157–68. Available from: <http://www.nature.com/doifinder/10.1038/onc.2013.178>
122. Hendrich B, Guy J, Ramsahoye B, Wilson VA, Bird A. Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev* [Internet]. Cold Spring Harbor Laboratory Press; 2001 Mar 15 [cited 2017 Oct 30];15(6):710–23. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11274056>
123. Monk M, Boubelik M, Lehnert S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* [Internet]. 1987 Mar [cited 2015 Jul 20];99(3):371–82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3653008>
124. Santos F, Hendrich B, Reik W, Dean W. Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev Biol* [Internet]. 2002 Jan 1 [cited 2015 Jul 13];241(1):172–82. Available from: <http://www.sciencedirect.com/science/article/pii/S0012160601905019>
125. Okano M, Bell DW, Haber D a, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* [Internet]. 1999 Oct 29;99(3):247–57. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10555141>
126. Erhardt S, Su I-H, Schneider R, Barton S, Bannister AJ, Perez-Burgos L, et al. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development* [Internet]. 2003 Sep [cited 2017 Oct 30];130(18):4235–48. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12900441>
127. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* [Internet]. 2006 Apr 21 [cited 2014 Mar 19];125(2):315–26. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867406003801>
128. Azuara V, Perry P, Sauer S, Spivakov M, Jørgensen HF, John RM, et al. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* [Internet]. Nature Publishing Group;

- 2006 May 29 [cited 2014 Jul 28];8(5):532–8. Available from: <http://www.nature.com/doi/10.1038/ncb1403>
129. Senner CE, Krueger F, Oxley D, Andrews S, Hemberger M. DNA methylation profiles define stem cell identity and reveal a tight embryonic-extraembryonic lineage boundary. *Stem Cells* [Internet]. 2012 Dec [cited 2013 Nov 1];30(12):2732–45. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23034951>
 130. Veillard A-C, Marks H, Bernardo AS, Jouneau L, Laloë D, Boulanger L, et al. Stable Methylation at Promoters Distinguishes Epiblast Stem Cells from Embryonic Stem Cells and the In Vivo Epiblasts. *Stem Cells Dev* [Internet]. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2014 Jun 12 [cited 2014 Aug 11];23(17):2014–29. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4142781&tool=pmcentrez&rendertype=abstract>
 131. Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, et al. Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013 Mar [cited 2013 Nov 16];20(3):311–6. Available from: <http://dx.doi.org/10.1038/nsmb.2510>
 132. Ficiz G, Hore TA, Santos F, Lee HJ, Dean W, Arand J, et al. FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency [Internet]. Vol. 13, *Cell Stem Cell*. 2013 [cited 2017 Aug 15]. Available from: <http://www.sciencedirect.com/science/article/pii/S1934590913002622>
 133. Habibi E, Brinkman AB, Arand J, Kroeze LI, Kerstens HDHD, Matarese F, et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* [Internet]. Elsevier; 2013 Sep 5 [cited 2014 Jul 23];13(3):360–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23850244>
 134. Marks H, Kalkan T, Menafrá R, Denissov S, Jones K, Hofemeister H, et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* [Internet]. 2012 Apr 27 [cited 2014 Jul 12];149(3):590–604. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398752&tool=pmcentrez&rendertype=abstract>
 135. Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, Ben-Yosef D, et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature* [Internet]. Nature Research; 2013 Oct 30 [cited 2017 Jun 27];504(7479):282–6. Available from: <http://www.nature.com/doi/10.1038/nature12745>
 136. Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, et al. The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Mol Cell* [Internet]. 2012 Dec [cited 2017 Jul 19];48(6):849–62. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S109727651200932X>
 137. Hirasawa R, Sasaki H. Dynamic transition of Dnmt3b expression in mouse pre- and early post-implantation embryos [Internet]. Vol. 9, *Gene Expression Patterns*. 2009 [cited 2017 Jul 19]. Available from: <http://www.sciencedirect.com/science/article/pii/S1567133X08001087>
 138. Benoist C, Chambon P. In vivo sequence requirements of the SV40 early promoter region. *Nature* [Internet]. 1981 Mar 26 [cited 2015 Aug 24];290(5804):304–10. Available from: <http://dx.doi.org/10.1038/290304a0>
 139. Gruss P, Dhar R, Khoury G. Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proc Natl Acad Sci U S A* [Internet]. 1981 Feb [cited 2015 Aug

- 24];78(2):943–7. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=319921&tool=pmcentrez&rendertype=abstract>
140. Schaffner W. Enhancers, enhancers – from their discovery to today’s universe of transcription enhancers. *Biol Chem* [Internet]. 2015 Jan 1 [cited 2017 Oct 23];396(4):311–27. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25719310>
 141. Banerji J, Olson L, Schaffner W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* [Internet]. Cell Press; 1983 Jul 1 [cited 2017 Oct 23];33(3):729–40. Available from: <http://www.sciencedirect.com/science/article/pii/0092867483900156?via%3Dihub>
 142. Gillies SD, Morrison SL, Oi VT, Tonegawa S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* [Internet]. Elsevier; 1983 Jul 1 [cited 2017 Oct 23];33(3):717–28. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6409417>
 143. Queen C, Baltimore D. Immunoglobulin gene transcription is activated by downstream sequence elements. *Cell* [Internet]. Cell Press; 1983 Jul 1 [cited 2017 Oct 23];33(3):741–8. Available from: <http://www.sciencedirect.com/science/article/pii/0092867483900168?via%3Dihub>
 144. Wang Y, Larsen AS, Peterlin BM. A tissue-specific transcriptional enhancer is found in the body of the HLA-DR alpha gene. *J Exp Med* [Internet]. 1987 Sep 1 [cited 2017 Oct 23];166(3):625–36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3476684>
 145. Shen RF, Li Y, Sifers RN, Wang H, Hardick C, Tsai SY, et al. Tissue-specific expression of the human alpha 1-antitrypsin gene is controlled by multiple cis-regulatory elements. *Nucleic Acids Res* [Internet]. 1987 Oct 26 [cited 2017 Oct 23];15(20):8399–415. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2823229>
 146. Delegeane AM, Ferland LH, Mellon PL. Tissue-specific enhancer of the human glycoprotein hormone alpha-subunit gene: dependence on cyclic AMP-inducible elements. *Mol Cell Biol* [Internet]. 1987 Nov [cited 2017 Oct 23];7(11):3994–4002. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2828923>
 147. Bouvagnet PF, Strehler EE, White GE, Strehler-Page MA, Nadal-Ginard B, Mahdavi V. Multiple positive and negative 5’ regulatory elements control the cell-type-specific expression of the embryonic skeletal myosin heavy-chain gene. *Mol Cell Biol* [Internet]. 1987 Dec [cited 2017 Oct 23];7(12):4377–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2830491>
 148. Jaynes JB, Johnson JE, Buskin JN, Gartside CL, Hauschka SD. The muscle creatine kinase gene is regulated by multiple upstream elements, including a muscle-specific enhancer. *Mol Cell Biol* [Internet]. 1988 Jan [cited 2017 Oct 23];8(1):62–70. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3336366>
 149. Harrison PR, Plumb M, Frampton J, Chambers I, Llewellyn D, Chester J, et al. cis and trans control of erythroid cell-specific gene expression during erythropoiesis. *J Cell Sci Suppl* [Internet]. 1988 [cited 2017 Oct 23];10:145–55. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3152055>
 150. Kruse F, Komro CT, Michnoff CH, MacDonald RJ. The cell-specific elastase I enhancer comprises two domains. *Mol Cell Biol* [Internet]. 1988 Feb [cited 2017 Oct 23];8(2):893–902. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3352608>
 151. Shelley CS, Baralle FE. Dual tissue-specific expression of apo-AII is directed by an upstream enhancer. *Nucleic Acids Res* [Internet]. 1987 May 11 [cited 2017 Oct

- 23];15(9):3801–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3035501>
152. Handy DE, Larsen SH, Karn RC, Hodes ME. Identification of a human salivary amylase gene. Partial sequence of genomic DNA suggests a mode of regulation different from that of mouse, Amy1. *Mol Biol Med* [Internet]. 1987 Jun [cited 2017 Oct 23];4(3):145–55. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2442579>
 153. Lagacé M, Howell BW, Burak R, Lusty CJ, Shore GC. Rat carbamyl-phosphate synthetase I gene. Promoter sequence and tissue-specific transcriptional regulation in vitro. *J Biol Chem* [Internet]. 1987 Aug 5 [cited 2017 Oct 23];262(22):10415–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3038878>
 154. Trainor CD, Stamler SJ, Engel JD. Erythroid-specific transcription of the chicken histone H5 gene is directed by a 3' enhancer. *Nature* [Internet]. 1987 Aug 27 [cited 2017 Oct 23];328(6133):827–30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3041221>
 155. Maxwell IH, Maxwell F, Glode LM. Regulated expression of a diphtheria toxin A-chain gene transfected into human cells: possible strategy for inducing cancer cell suicide. *Cancer Res* [Internet]. 1986 Sep [cited 2017 Oct 23];46(9):4660–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3460697>
 156. Rogers BL, Sobnosky MG, Saunders GF. Transcriptional enhancer within the human placental lactogen and growth hormone multigene cluster. *Nucleic Acids Res* [Internet]. 1986 Oct 10 [cited 2017 Oct 23];14(19):7647–59. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3774541>
 157. Choi O-R, Engel JD. A 3' enhancer is required for temporal and tissue-specific transcriptional activation of the chicken adult β -globin gene. *Nature* [Internet]. 1986 Oct 23 [cited 2017 Oct 23];323(6090):731–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3022151>
 158. Widen SG, Papaconstantinou J. Liver-specific expression of the mouse alpha-fetoprotein gene is mediated by cis-acting DNA elements. *Proc Natl Acad Sci U S A* [Internet]. 1986 Nov [cited 2017 Oct 23];83(21):8196–200. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2430280>
 159. Buckingham M, Alonso S, Barton P, Cohen A, Daubas P, Garner I, et al. Actin and myosin multigene families. Their expression during the formation and maturation of striated muscle. *Am J Med Genet* [Internet]. 1986 Dec [cited 2017 Oct 23];25(4):623–34. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3789022>
 160. Rochford R, Campbell BA, Villarreal LP. A pancreas specificity results from the combination of polyomavirus and Moloney murine leukemia virus enhancer. *Proc Natl Acad Sci U S A* [Internet]. 1987 Jan [cited 2017 Oct 23];84(2):449–53. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3025873>
 161. Hammer RE, Krumlauf R, Camper SA, Brinster RL, Tilghman SM. Diversity of alpha-fetoprotein gene expression in mice is generated by a combination of separate enhancer elements. *Science* [Internet]. 1987 Jan 2 [cited 2017 Oct 23];235(4784):53–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2432657>
 162. Kohno K, Sullivan M, Yamada Y. Structure of the promoter of the rat type II procollagen gene. *J Biol Chem* [Internet]. 1985 Apr 10 [cited 2017 Oct 23];260(7):4441–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2984204>
 163. Mosthaf L, Pawlita M, Gruss P. A viral enhancer element specifically active in human haematopoietic cells. *Nature* [Internet]. [cited 2017 Oct 23];315(6020):597–600. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2989696>

164. Okazaki K, Yasuda K, Kondoh H, Okada TS. DNA sequences responsible for tissue-specific expression of a chicken alpha-crystallin gene in mouse lens cells. *EMBO J* [Internet]. European Molecular Biology Organization; 1985 Oct [cited 2017 Oct 23];4(10):2589–95. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/4054099>
165. Wetterstrand KA. DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI) [Internet]. [cited 2017 Oct 23]. Available from: <https://www.genome.gov/sequencingcostsdata/>
166. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, et al. High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. Hannehalli S, editor. *PLoS One* [Internet]. Public Library of Science; 2010 Feb 9 [cited 2017 Oct 23];5(2):e9129. Available from: <http://dx.plos.org/10.1371/journal.pone.0009129>
167. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 Sep 7 [cited 2014 Jul 10];13(9):613–26. Available from: <http://dx.doi.org/10.1038/nrg3207>
168. Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* [Internet]. Cell Press; 2016 Nov 17 [cited 2017 Oct 23];167(5):1170–87. Available from: <http://www.sciencedirect.com/science/article/pii/S009286741631251X?via%3Dihub>
169. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* [Internet]. Cold Spring Harbor Laboratory Press; 2011 Nov 1 [cited 2017 Oct 23];25(21):2227–41. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22056668>
170. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* [Internet]. Elsevier; 2006 Aug 25 [cited 2014 Jul 9];126(4):663–76. Available from: <http://www.cell.com/article/S0092867406009767/fulltext>
171. Thanos D, Maniatis T. Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell* [Internet]. Cell Press; 1995 Dec 29 [cited 2017 Oct 23];83(7):1091–100. Available from: <http://www.sciencedirect.com/science/article/pii/0092867495901361>
172. Guturu H, Doxey AC, Wenger AM, Bejerano G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos Trans R Soc Lond B Biol Sci* [Internet]. The Royal Society; 2013 Dec 19 [cited 2017 Oct 23];368(1632):20130029. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24218641>
173. Kulkarni MM, Arnosti DN. Information display by transcriptional enhancers. *Development* [Internet]. The Company of Biologists Ltd; 2003 Dec 29 [cited 2017 Oct 23];130(26):6569–75. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14660545>
174. Arnosti DN, Kulkarni MM. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* [Internet]. Wiley Subscription Services, Inc., A Wiley Company; 2005 Apr 1 [cited 2017 Oct 23];94(5):890–8. Available from: <http://doi.wiley.com/10.1002/jcb.20352>
175. Weake VM, Workman JL. Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2010 Jun 27 [cited 2017 Oct 23];11(6):426–37. Available from: <http://www.nature.com/doifinder/10.1038/nrg2781>
176. Roeder RG. Transcriptional regulation and the role of diverse coactivators in animal

- cells. *FEBS Lett* [Internet]. 2005 Feb 7 [cited 2017 Oct 23];579(4):909–15. Available from: <http://doi.wiley.com/10.1016/j.febslet.2004.12.007>
177. Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, et al. Functional annotation of native enhancers with a Cas9–histone demethylase fusion. *Nat Methods* [Internet]. Nature Research; 2015 Mar 16 [cited 2017 Oct 23];12(5):401–3. Available from: <http://www.nature.com/doi/10.1038/nmeth.3325>
 178. Jeronimo C, Robert F. The Mediator Complex: At the Nexus of RNA Polymerase II Transcription. *Trends Cell Biol* [Internet]. Elsevier Current Trends; 2017 Oct 1 [cited 2017 Oct 23];27(10):765–83. Available from: <http://www.sciencedirect.com/science/article/pii/S0962892417301162?via%3Dihub>
 179. Whyte WAA, Orlando DAA, Hnisz D, Abraham BJJ, Lin CYY, Kagey MHH, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* [Internet]. Elsevier Inc.; 2013 Apr 11 [cited 2013 Apr 16];153(2):307–19. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867413003929>
 180. Kagey MHM, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* [Internet]. 2010 Sep 23 [cited 2013 May 3];467(7314):430–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20720539>
 181. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 May 17 [cited 2013 Oct 17];485(7398):376–80. Available from: <http://www.nature.com/doi/10.1038/nature11082>
 182. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* [Internet]. Nature Research; 2012 Apr 11 [cited 2017 Oct 23];485(7398):381–5. Available from: <http://www.nature.com/doi/10.1038/nature11049>
 183. Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, et al. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* [Internet]. EMBO Press; 2013 Dec 11 [cited 2017 Oct 31];32(24):3119–29. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24185899>
 184. Downen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, et al. Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell* [Internet]. 2014 Oct [cited 2017 Oct 31];159(2):374–87. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867414011799>
 185. Alipour E, Marko JF. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* [Internet]. 2012 Dec 1 [cited 2017 Oct 31];40(22):11202–12. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23074191>
 186. Goloborodko A, Marko JF, Mirny LA. Chromosome Compaction by Active Loop Extrusion. *Biophys J* [Internet]. 2016 May 24 [cited 2017 Oct 31];110(10):2162–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27224481>
 187. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* [Internet]. 2016 May 31 [cited 2017 Oct 31];15(9):2038–49. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27210764>
 188. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* [Internet]. Elsevier; 2015 Mar 3 [cited 2017 Oct 31];10(8):1297–

309. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25732821>
189. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Jul 2 [cited 2014 Jul 9];advance on. Available from: <http://dx.doi.org/10.1038/nature13417>
 190. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013 Nov 14 [cited 2014 Jul 10];503(7475):290–4. Available from: <http://dx.doi.org/10.1038/nature12644>
 191. de Laat W, Dekker J. 3C-based technologies to study the shape of the genome. *Methods* [Internet]. 2012 Nov [cited 2014 Jul 23];58(3):189–91. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874834&tool=pmcentrez&rendertype=abstract>
 192. Cannavò E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, et al. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr Biol* [Internet]. Cell Press; 2016 Jan 11 [cited 2017 Oct 23];26(1):38–51. Available from: <http://www.sciencedirect.com/science/article/pii/S0960982215014281?via%3Dihub>
 193. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, et al. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet* [Internet]. Nature Research; 2011 Mar 20 [cited 2017 Oct 23];43(4):379–86. Available from: <http://www.nature.com/doifinder/10.1038/ng.790>
 194. Loven J, Hoke HAA, Lin CYY, Lau A, Orlando DAA, Vakoc CRR, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* [Internet]. United States: Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA.; 2013 Apr 11 [cited 2013 Apr 16];153(2):320–34. Available from: <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medc&NEWS=N&AN=23582323>
 195. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell* [Internet]. 2013 [cited 2013 Nov 4]; Available from: <http://www.sciencedirect.com/science/article/pii/S0092867413012270>
 196. Gluzman Y. SV40-transformed simian cells support the replication of early SV40 mutants. *Cell* [Internet]. 1981 Jan [cited 2017 Apr 11];23(1):175–82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6260373>
 197. Smith AG, Hooper ML. Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells. *Dev Biol* [Internet]. 1987 May [cited 2016 Feb 20];121(1):1–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3569655>
 198. Jouneau A, Ciaudo C, Sismeiro O, Brochard V, Jouneau L, Vandormael-Pournin S, et al. Naive and primed murine pluripotent stem cells have distinct miRNA expression profiles. *RNA* [Internet]. Cold Spring Harbor Laboratory Press; 2012 Feb [cited 2017 Apr 11];18(2):253–64. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22201644>
 199. Chen J, Nathans J. Estrogen-Related Receptor β /NR3B2 Controls Epithelial Cell Fate and Endolymph Production by the Stria Vascularis. *Dev Cell* [Internet]. 2007 [cited 2017 Apr 11];13(3):325–37. Available from:

- <http://www.sciencedirect.com/science/article/pii/S1534580707002729>
200. Festuccia N, Osorno R, Halbritter F, Karwacki-Neisius V, Navarro P, Colby D, et al. Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell* [Internet]. Elsevier; 2012 Oct 5 [cited 2013 Dec 11];11(4):477–90. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23040477>
 201. Chambers I, Silva J, Colby D, Nichols J, Nijmeijer B, Robertson M, et al. Nanog safeguards pluripotency and mediates germline development. *Nature* [Internet]. Nature Publishing Group; 2007 Dec 20 [cited 2017 Apr 11];450(7173):1230–4. Available from: <http://www.nature.com/doifinder/10.1038/nature06403>
 202. Vanacker JM, Pettersson K, Gustafsson JÅ, Laudet V. Transcriptional targets shared by estrogen receptor-related receptors (ERRs) and estrogen receptor (ER) α , but not by ER β . *EMBO J* [Internet]. 1999 Aug 2 [cited 2017 Apr 11];18(15):4270–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10428965>
 203. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20:307–15.
 204. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* [Internet]. Oxford University Press; 2010 Mar 15 [cited 2017 Apr 13];26(6):841–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20110278>
 205. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* [Internet]. BioMed Central; 2009 Jan 4 [cited 2014 Jul 9];10(3):R25. Available from: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25>
 206. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* [Internet]. Nature Research; 2012 Mar 4 [cited 2017 Apr 13];9(4):357–9. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.1923>
 207. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. [cited 2016 May 20]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
 208. NCBI. NCBI Sequence Read Archive [Internet]. [cited 2016 May 20]. Available from: <http://www.ncbi.nlm.nih.gov/Traces/sra>
 209. FileZilla. FileZilla - The free FTP solution [Internet]. 2013 [cited 2017 Apr 14]. Available from: <https://filezilla-project.org/>
 210. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol* [Internet]. Public Library of Science; 2013 Jan 8 [cited 2015 Jun 14];9(8):e1003118. Available from: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118>
 211. The Gimp Team. GNU Image Manipulation Program (GIMP) [Internet]. Available from: <https://www.gimp.org/>
 212. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* [Internet]. 2010 May 28 [cited 2014 Jul 10];38(4):576–89. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1097276510003667>
 213. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* [Internet]. 2015 Apr 20 [cited 2015 Jan 21];43(7):e47–e47. Available from:

- <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv007>
214. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. 2008;24:1547–8.
 215. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* [Internet]. 2008 Jan [cited 2013 Nov 6];9(9):R137. Available from: <http://genomebiology.com/2008/9/9/R137>
 216. The Mendeley Team. Mendeley [Internet]. [cited 2017 Apr 14]. Available from: <http://www.mendeley.com/>
 217. Microsoft. Microsoft Office 2013.
 218. Broad Institute. Picard tools [Internet]. <https://broadinstitute.github.io/picard/>. 2016 [cited 2017 Apr 14]. Available from: <http://broadinstitute.github.io/picard/>
 219. The Putty Team. PuTTY [Internet]. [cited 2017 Apr 14]. Available from: <http://www.chiark.greenend.org.uk/>
 220. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. Vienna, Austria; 2016 [cited 2017 Apr 14]. Available from: <https://www.r-project.org/>
 221. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009 Aug 15 [cited 2017 Apr 14];25(16):2078–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19505943>
 222. Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* [Internet]. 2009 Oct 1 [cited 2017 Apr 14];25(19):2607–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19654119>
 223. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* [Internet]. 2014 Apr 28 [cited 2014 Jul 9];30(15):2114–20. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4103590&tool=pmcentrez&rendertype=abstract>
 224. Lang D. XML: Tools for parsing and generating XML within R and S-Plus [Internet]. R package version. 2012 [cited 2017 Apr 14]. Available from: <https://cran.r-project.org/package=XML>
 225. Tosolini M, Jouneau A. From Naive to Primed Pluripotency: In Vitro Conversion of Mouse Embryonic Stem Cells in Epiblast Stem Cells. In Humana Press, New York, NY; 2015 [cited 2017 Nov 7]. p. 209–16. Available from: http://link.springer.com/10.1007/7651_2015_208
 226. Livak KJ, Schmittgen TD. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2- $\Delta\Delta$ CT Method. *Methods* [Internet]. 2001 Dec [cited 2017 Apr 25];25(4):402–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11846609>
 227. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* [Internet]. Oxford University Press; 2001 May 1 [cited 2017 Apr 25];29(9):45e–45. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11328886>
 228. Tomaz RA. The Role of the Histone Demethylase Jmjd2c/Kdm4c in Mouse Embryonic Stem Cell Pluripotency. Imperial College London; 2016.
 229. The Bioconductor Dev Team. BSgenome.Mmusculus.UCSC.mm9: Full genome sequences for Mus musculus (UCSC version mm9). R package version 1.4.0. 2014.

230. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* [Internet]. 2006 Jan 31 [cited 2014 Jan 20];103(5):1412–7. Available from: <http://www.pnas.org/content/103/5/1412.full>
231. Edgar R, Lash A. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository. *NCBI Handb.* 2002;0.
232. Smyth GK. Limma: Linear Models for Microarray Data. *Bioinformatics.* 2005;pages:397–420.
233. Dunning M, Barbosa-Morais N. Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* [Internet]. 2008 [cited 2013 Aug 14];9(85). Available from: <http://www.biomedcentral.com/1471-2105/9/85/>
234. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* [Internet]. 2004 Aug 27 [cited 2013 Aug 8];573(1–3):83–92. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15327980>
235. Hong F, Wittner B, Breitling R, Smith C, Battke F. RankProd: Rank Product method for identifying differentially expressed genes with application in meta-analysis. R package version 2.30.0; 2011.
236. Fisher RA. Statistical methods for research workers. Vol. 354, Biological monographs and manuals. Edinburgh: Oliver and Boyd; 1925. xv, 356 .
237. (US) NC for BI. Using the SRA Toolkit to convert .sra files into other formats [Internet]. National Center for Biotechnology Information (US); 2011 [cited 2014 Aug 4]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK158900/>
238. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* [Internet]. Oxford University Press; 2010 Sep 1 [cited 2014 Jan 21];26(17):2204–7. Available from: <http://bioinformatics.oxfordjournals.org/content/26/17/2204.short>
239. James Kent W, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2002 Jun [cited 2017 Apr 26];12(6):996–1006. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12045153>
240. Humburg P, Bulger D, Stone G. Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics.* 2008;(9):343.
241. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 Sep [cited 2013 Oct 17];7(9):1728–40. Available from: <http://dx.doi.org/10.1038/nprot.2012.101>
242. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* [Internet]. Oxford University Press; 2014 Jan 15 [cited 2016 Sep 2];30(2):284–6. Available from: <http://www.bioconductor.org>.
243. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* [Internet]. Nature Research; 2013 Jul 18 [cited 2017 Apr 26];8(8):1551–66. Available from: <http://www.nature.com/doifinder/10.1038/nprot.2013.092>
244. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R.

- ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* [Internet]. Oxford University Press; 2011 Jan [cited 2015 Aug 28];39(Database issue):D712-7. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013724&tool=pmcentrez&rendertype=abstract>
245. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB - A database for integrating human functional interaction networks. *Nucleic Acids Res* [Internet]. Oxford University Press; 2009 Jan [cited 2017 Apr 26];37(SUPPL. 1):D623-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18940869>
 246. Hastie AT, Tibshirani R, Narasimhan B, Chu G. impute: Imputation for microarray data. R pack. R package version 1.32.0;
 247. Chambers I, Tomlinson SR. The transcriptional foundation of pluripotency. *Development* [Internet]. 2009 [cited 2017 Jul 11];136(14). Available from: <http://dev.biologists.org/content/136/14/2311.long>
 248. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, et al. Dissecting self-renewal in stem cells with RNA interference. *Nature* [Internet]. 2006 Aug 3 [cited 2013 Nov 8];442(7102):533–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16767105>
 249. van den Berg DLC, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, et al. An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* [Internet]. 2010 Apr 2 [cited 2014 May 23];6(4):369–81. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2860243&tool=pmcentrez&rendertype=abstract>
 250. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* [Internet]. 2007 May [cited 2017 Jul 18];129(4):823–37. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867407006009>
 251. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* [Internet]. National Academy of Sciences; 2010 Dec 14 [cited 2013 Sep 23];107(50):21931–6. Available from: <http://www.pnas.org/cgi/content/long/107/50/21931>
 252. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* [Internet]. NIH Public Access; 2008 [cited 2017 Jul 20];40(7):897. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769248/>
 253. Müller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, et al. Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. *Cell* [Internet]. 2002 [cited 2017 Jul 20];111(2):197–208. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867402009765>
 254. Koh KP, Yabuuchi A, Rao S, Huang Y, Cunliffe K, Nardone J, et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* [Internet]. NIH Public Access; 2011 Feb 4 [cited 2017 Jun 29];8(2):200–13. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21295276>
 255. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol* [Internet]. 2011 Jun 20 [cited 2017 Jul 18];12(6):R54. Available from:

- <http://www.ncbi.nlm.nih.gov/pubmed/21689397>
256. Ficiz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* [Internet]. 2011 Apr 3 [cited 2017 Jul 18];473(7347):398–402. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21460836>
 257. Singer ZSS, Yong J, Tischler J, Hackett JAA, Altinok A, Surani MAA, et al. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Mol Cell* [Internet]. 2014 Jul 17 [cited 2014 Jul 18];55(2):319–31. Available from: <http://www.sciencedirect.com/science/article/pii/S1097276514005632>
 258. Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* [Internet]. Cold Spring Harbor Laboratory Press; 2011 Apr 1 [cited 2015 Jan 8];25(7):679–84. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21460036>
 259. Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* [Internet]. *Nature Research*; 2011 May 8 [cited 2017 Jun 27];473(7347):394–7. Available from: <http://www.nature.com/doifinder/10.1038/nature10102>
 260. Szulwach KE, Li X, Li Y, Song C-X, Han JW, Kim S, et al. Integrating 5-Hydroxymethylcytosine into the Epigenomic Landscape of Human Embryonic Stem Cells. Pearson CE, editor. *PLoS Genet* [Internet]. 679-684; 2011 Jun 23 [cited 2017 Jun 27];7(6):e1002154. Available from: <http://dx.plos.org/10.1371/journal.pgen.1002154>
 261. Huang Y, Chavez L, Chang X, Wang X, Pastor WA, Kang J, et al. Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 2014 Jan 28 [cited 2017 Jul 19];111(4):1361–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24474761>
 262. Najm FJJ, Chenoweth JGG, Anderson PDD, Nadeau JHH, Redline RWW, McKay RDGDG, et al. Isolation of epiblast stem cells from preimplantation mouse embryos. *Cell Stem Cell* [Internet]. 2011 [cited 2017 Jul 20];8(3):318–25. Available from: <http://www.sciencedirect.com/science/article/pii/S193459091100049X>
 263. Iwafuchi-Doi M, Matsuda K, Murakami K, Niwa H, Tesar PJ, Aruga J, et al. Transcriptional regulatory networks in epiblast cells and during anterior neural plate development as modeled in epiblast stem cells. *Development* [Internet]. 2012 Nov 1 [cited 2017 Jul 20];139(21):3926–37. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22992956>
 264. Hayashi K, Ohta H, Kurimoto K, Aramaki S, Saitou M. Reconstitution of the Mouse Germ Cell Specification Pathway in Culture by Pluripotent Stem Cells. *Cell* [Internet]. 2011 Aug 19 [cited 2017 Jul 20];146(4):519–32. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21820164>
 265. Pan J-B, Hu S-C, Shi D, Cai M-C, Li Y-B, Zou Q, et al. PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. Palsson A, editor. *PLoS One* [Internet]. Public Library of Science; 2013 Jan [cited 2014 Aug 4];8(12):e80747. Available from: <http://dx.plos.org/10.1371/journal.pone.0080747>
 266. Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* [Internet]. *Nature Research*; 2012 Mar 28 [cited 2017 Aug 15];484(7394):339–44. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22456710>

267. von Meyenn F, Iurlaro M, Habibi E, Liu NQ, Salehzadeh-Yazdi A, Santos F, et al. Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. *Mol Cell* [Internet]. Elsevier; 2016 Jun 16 [cited 2017 Aug 14];62(6):848–61. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1097276516301010>
268. Babraham Bioinformatics. Bismark Bisulfite Mapper User Guide v0.13.0 [Internet]. 2014 [cited 2014 Nov 10]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/bismark/Bismark_User_Guide.pdf
269. Tanaka S, Kunath T, Hadjantonakis A-K, Nagy A, Rossant J. Promotion of Trophoblast Stem Cell Proliferation by FGF4. *Science* (80-) [Internet]. 1998 [cited 2017 Aug 15];282(5396). Available from: <http://science.sciencemag.org/content/282/5396/2072.full>
270. Kunath T, Arnaud D, Uy GD, Okamoto I, Chureau C, Yamanaka Y, et al. Imprinted X-inactivation in extra-embryonic endoderm cell lines from mouse blastocysts. *Development* [Internet]. 2005 [cited 2017 Aug 15];132(7). Available from: <http://dev.biologists.org/content/132/7/1649>
271. Rossant J, Ofer L. Properties of extra-embryonic ectoderm isolated from postimplantation mouse embryos. *Development* [Internet]. 1977 [cited 2017 Aug 15];39(1). Available from: <http://dev.biologists.org/content/39/1/183.long>
272. Cabezas-Wallscheid N, Klimmeck D, Hansson J, Lipka DB, Reyes A, Wang Q, et al. Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. *Cell Stem Cell* [Internet]. 2014 Oct 2 [cited 2017 Aug 18];15(4):507–22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25158935>
273. Yu W, McIntosh C, Lister R, Zhu I, Han Y, Ren J, et al. Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. *Genome Res* [Internet]. 2014 Oct [cited 2017 Aug 18];24(10):1613–23. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25170028>
274. Yu W, Briones V, Lister R, McIntosh C, Han Y, Lee EY, et al. CG hypomethylation in Lsh^{-/-} mouse embryonic fibroblasts is associated with de novo H3K4me1 formation and altered cellular plasticity. *Proc Natl Acad Sci* [Internet]. 2014 Apr 22 [cited 2017 Aug 18];111(16):5890–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24711395>
275. Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, et al. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet* [Internet]. Nature Research; 2011 Jun 26 [cited 2017 Aug 18];43(8):811–4. Available from: <http://www.nature.com/doifinder/10.1038/ng.864>
276. Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Mol Syst Biol* [Internet]. 2014 Apr 16 [cited 2017 Apr 27];7(1):562–562. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22186736>
277. Song C-XX, Szulwach KE, Dai Q, Fu Y, Mao S-QQ, Lin L, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* [Internet]. 2013 Apr 25 [cited 2017 Apr 27];153(3):678–91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23602153>
278. Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, et al. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* [Internet]. Nature Research; 2011 Mar 30 [cited 2017 Apr 27];473(7347):389–93. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21451524>

279. Chen Q, Chen Y, Bian C, Fujiki R, Yu X. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature* [Internet]. 2012 Dec 9 [cited 2017 Apr 27];493(7433):561–4. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23222540>
280. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* [Internet]. 2015 Jan 21 [cited 2017 Aug 18];520(7546):243–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25607372>
281. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* [Internet]. Nature Publishing Group; 2007 Mar 4 [cited 2017 Aug 19];39(3):311–8. Available from: <http://www.nature.com/doi/10.1038/ng1966>
282. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* [Internet]. 2014 Nov 19 [cited 2017 Apr 27];515(7527):355–64. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25409824>
283. Factor DCC, Corradin O, Zentner GEE, Saiakhova A, Song L, Chenoweth JGG, et al. Epigenomic Comparison Reveals Activation of “Seed” Enhancers during Transition from Naive to Primed Pluripotency. *Cell Stem Cell* [Internet]. Elsevier; 2014 Jun 5 [cited 2014 Jun 5];14(6):854–63. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24905169>
284. Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I, et al. Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Res* [Internet]. 2010 Jul [cited 2017 Aug 19];38(13):4246–53. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20223770>
285. Otani J, Nankumo T, Arita K, Inamoto S, Ariyoshi M, Shirakawa M. Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain. *EMBO Rep* [Internet]. 2009 Nov 16 [cited 2017 Aug 19];10(11):1235–41. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19834512>
286. Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* [Internet]. 2007 Aug 9 [cited 2017 Aug 19];448(7154):714–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17687327>
287. Noh K-M, Wang H, Kim HR, Wenderski W, Fang F, Li CH, et al. Engineering of a Histone-Recognition Domain in Dnmt3a Alters the Epigenetic Landscape and Phenotypic Features of Mouse ESCs. *Mol Cell* [Internet]. 2015 Jul [cited 2017 Aug 19];59(1):89–103. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1097276515003536>
288. King AD, Huang K, Rubbi L, Liu S, Wang C-Y, Wang Y, et al. Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells. *Cell Rep* [Internet]. 2016 Sep [cited 2017 Aug 19];17(1):289–302. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S2211124716311792>
289. Soufi A, Donahue G, Zaret KS. Facilitators and Impediments of the Pluripotency Reprogramming Factors’ Initial Engagement with the Genome. *Cell* [Internet]. 2012 [cited 2017 Aug 21];151(5):994–1004. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867412012986>
290. Yang S-H, Kalkan T, Morissroe C, Marks H, Stunnenberg H, Smith A, et al. Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. *Cell Rep* [Internet]. 2014 Jun 26 [cited 2015 May 19];7(6):1968–81.

Available from: <http://linkinghub.elsevier.com/retrieve/pii/S2211124714004318>

291. Chen J, Chen X, Li M, Liu X, Gao Y, Kou X, et al. Hierarchical Oct4 Binding in Concert with Primed Epigenetic Rearrangements during Somatic Cell Reprogramming. *Cell Rep* [Internet]. 2016 Feb 16 [cited 2017 Aug 21];14(6):1540–54. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26832419>
292. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS, et al. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* [Internet]. Elsevier; 2002 Feb 1 [cited 2017 Aug 21];9(2):279–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11864602>
293. Pasini D, Malatesta M, Jung HR, Walfridsson J, Willer A, Olsson L, et al. Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. *Nucleic Acids Res* [Internet]. Oxford University Press; 2010 Aug [cited 2017 Aug 21];38(15):4958–69. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20385584>
294. Giresi PPG, Kim J, McDaniel RMR, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2007 Jun [cited 2013 Aug 14];17(6):877–85. Available from: <http://genome.cshlp.org/content/17/6/877.short>
295. Consortium TGO. Gene Ontology Consortium: going forward. *Nucleic Acids Res* [Internet]. Riva del Garda, Italy; 2015 Jan 28 [cited 2017 Aug 28];43(D1):D1049–56. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1179>
296. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* [Internet]. Nature Publishing Group; 2000 May 1 [cited 2017 Aug 28];25(1):25–9. Available from: <http://www.nature.com/doifinder/10.1038/75556>
297. Nishiyama A, Sharov AA, Piao Y, Amano M, Amano T, Hoang HG, et al. Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Sci Rep* [Internet]. 2013 Dec 6 [cited 2017 May 2];3(1):1390. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23462645>
298. Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M. Microarray experiments and factors which affect their reliability. *Biol Direct* [Internet]. BioMed Central; 2015 Sep 3 [cited 2017 Oct 2];10:46. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26335588>
299. Ye S, Li P, Tong C, Ying Q-L. Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. *EMBO J* [Internet]. 2013 Aug 13 [cited 2017 Oct 2];32(19):2548–60. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23942238>
300. Yuan K, Ai W-B, Wan L-Y, Tan X, Wu J-F. The miR-290-295 cluster as multi-faceted players in mouse embryonic stem cells. *Cell Biosci* [Internet]. BioMed Central; 2017 Dec 7 [cited 2017 Oct 2];7(1):38. Available from: <http://cellandbioscience.biomedcentral.com/articles/10.1186/s13578-017-0166-2>
301. Percharde M, Laval F, Ng J-H, Kumar V, Tomaz RA, Martin N, et al. Nco3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. *Genes Dev* [Internet]. 2012 Oct 15 [cited 2013 Oct 21];26(20):2286–98. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3475801&tool=pmcentrez&rendertype=abstract>

302. Wu Z, Yang M, Liu H, Guo H, Wang Y, Cheng H, et al. Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. *J Biol Chem* [Internet]. 2012 Nov 2 [cited 2013 Dec 11];287(45):38295–304. Available from: <http://www.jbc.org/content/287/45/38295.full>
303. Jiang J, Chan Y-S, Loh Y-H, Cai J, Tong G-Q, Lim C-A, et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* [Internet]. 2008;10(3):353–60. Available from: <http://www.nature.com/doi/10.1038/ncb1698>
304. Ren G, Jin W, Cui K, Rodriguez J, Hu G, Zhang Z, et al. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol Cell* [Internet]. 2017 Sep 21 [cited 2017 Oct 3];67(6):1049–1058.e6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28938092>
305. Martello G, Sugimoto T, Diamanti E, Joshi A, Hannah R, Ohtsuka S, et al. Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. *Cell Stem Cell* [Internet]. 2012 Oct 5 [cited 2013 Oct 23];11(4):491–504. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3465555&tool=pmcentrez&rendertype=abstract>
306. Zhang X, Zhang J, Wang T, Esteban MA, Pei D. Esrrb Activates Oct4 Transcription and Sustains Self-renewal and Pluripotency in Embryonic Stem Cells. *J Biol Chem* [Internet]. 2008 Dec 19 [cited 2013 Dec 11];283(51):35825–33. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18957414>
307. Feng B, Jiang J, Kraus P, Ng J-H, Heng J-CD, Chan Y-S, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* [Internet]. Nature Publishing Group; 2009 Feb 11 [cited 2013 Sep 21];11(2):197–203. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19136965>
308. Xie C-Q, Jeong Y, Fu M, Bookout AL, Garcia-Barrio MT, Sun T, et al. Expression Profiling of Nuclear Receptors in Human and Mouse Embryonic Stem Cells. *Mol Endocrinol* [Internet]. 2009 May [cited 2015 Sep 21];23(5):724–33. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2675959&tool=pmcentrez&rendertype=abstract>
309. Tseng GGC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res* [Internet]. 2012 May [cited 2013 Aug 14];40(9):3785–99. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3351145&tool=pmcentrez&rendertype=abstract>
310. Acampora D, Di Giovannantonio LG, Simeone A. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development* [Internet]. 2013;140(1):43–55. Available from: <http://dev.biologists.org/cgi/doi/10.1242/dev.085290>
311. Xiao S, Lu J, Sridhar B, Cao X, Yu P, Zhao T, et al. SMARCD1 Contributes to the Regulation of Naive Pluripotency by Interacting with Histone Citrullination. *Cell Rep*. 2017;18(13):3117–28.
312. Choi HW, Joo JY, Hong YJ, Kim JS, Song H, Lee JW, et al. Distinct Enhancer Activity of Oct4 in Naive and Primed Mouse Pluripotency. *Stem Cell Reports*. 2016;7(5):911–26.
313. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* [Internet]. Nature Research; 2012 Jul 1 [cited 2017 Apr 27];488(7409):116–20. Available from: <http://www.nature.com/doi/10.1038/nature11243>

314. Rubtsov MA, Polikanov YS, Bondarenko VA, Wang Y-H, Studitsky VM. Chromatin structure can strongly facilitate enhancer action over a distance. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 2006 Nov 21 [cited 2017 Oct 14];103(47):17690–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17101994>
315. Feng Z-X, Li Q-Z. Recognition of long-range enhancer-promoter interactions by adding genomic signatures of segmented regulatory regions. *Genomics* [Internet]. 2017 Jun [cited 2017 Oct 14]; Available from: <http://linkinghub.elsevier.com/retrieve/pii/S088875431730040X>
316. Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* [Internet]. 2012 Feb 1 [cited 2017 Apr 27];482(7384):221–5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22297846>
317. Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, et al. Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks. Reik W, editor. *PLoS Genet* [Internet]. Public Library of Science; 2012 Jan 5 [cited 2017 Aug 15];8(1):e1002440. Available from: <http://dx.plos.org/10.1371/journal.pgen.1002440>
318. Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, et al. The DNA methylation landscape of human early embryos. *Nature* [Internet]. *Nature Research*; 2014 Jul 23 [cited 2017 Oct 17];511(7511):606–10. Available from: <http://www.nature.com/doifinder/10.1038/nature13544>
319. Sanchez-Delgado M, Court F, Vidal E, Medrano J, Monteagudo-Sánchez A, Martín-Trujillo A, et al. Human Oocyte-Derived Methylation Differences Persist in the Placenta Revealing Widespread Transient Imprinting. Bartolomei MS, editor. *PLOS Genet* [Internet]. Public Library of Science; 2016 Nov 11 [cited 2017 Oct 17];12(11):e1006427. Available from: <http://dx.plos.org/10.1371/journal.pgen.1006427>
320. Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, et al. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2017 Feb 1 [cited 2017 Oct 14];27(2):246–58. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27895109>
321. Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet* [Internet]. *Nature Research*; 2017 May 30 [cited 2017 Jul 20];18(9):517–34. Available from: <http://www.nature.com/doifinder/10.1038/nrg.2017.33>
322. Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, et al. Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Mol Cell* [Internet]. 2011 May 20 [cited 2017 Apr 27];42(4):451–64. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1097276511002838>
323. Williams K, Christensen J, Pedersen MT, Johansen J V., Cloos PAC, Rappsilber J, et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* [Internet]. *Nature Research*; 2011 Apr 13 [cited 2017 Oct 18];473(7347):343–8. Available from: <http://www.nature.com/doifinder/10.1038/nature10066>
324. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D. Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions. Barsh GS, editor. *PLoS Genet* [Internet]. Public Library of Science; 2013 Dec 19 [cited 2017 Oct 19];9(12):e1003994. Available from: <http://dx.plos.org/10.1371/journal.pgen.1003994>

325. Dubois-Chevalier J, Oger F, Dehondt H, Firmin FF, Gheeraert C, Staels B, et al. A dynamic CTCF chromatin binding landscape promotes DNA hydroxymethylation and transcriptional induction of adipocyte differentiation. *Nucleic Acids Res* [Internet]. Oxford University Press; 2014 Sep 29 [cited 2017 Oct 18];42(17):10943–59. Available from: <http://academic.oup.com/nar/article/42/17/10943/2902698/A-dynamic-CTCF-chromatin-binding-landscape>
326. Dawlaty MM, Breiling A, Le T, Barrasa MI, Raddatz G, Gao Q, et al. Loss of Tet Enzymes Compromises Proper Differentiation of Embryonic Stem Cells. *Dev Cell* [Internet]. 2014 Apr [cited 2017 Oct 19];29(1):102–11. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1534580714001555>
327. Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, Kim A, et al. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* [Internet]. 2012 Jun [cited 2017 Oct 19];149(6):1368–80. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S009286741200534X>
328. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* [Internet]. American Association for the Advancement of Science; 2012 May 18 [cited 2017 Oct 19];336(6083):934–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22539555>
329. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, et al. Editing DNA Methylation in the Mammalian Genome. *Cell* [Internet]. 2016 Sep [cited 2017 Oct 19];167(1):233–247.e17. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0092867416311539>
330. Li M, Belmonte JCI. Ground rules of the pluripotency gene regulatory network. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2017 Jan 3 [cited 2017 Oct 19];18(3):180–91. Available from: <http://www.nature.com/doifinder/10.1038/nrg.2016.156>
331. Kubo N, Ishii H, Gorkin D, Meitinger F, Xiong X, Fang R, et al. Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. *doi.org* [Internet]. Cold Spring Harbor Laboratory; 2017 Mar 20 [cited 2017 Oct 21];118737. Available from: <https://www.biorxiv.org/content/early/2017/03/20/118737>
332. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, et al. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell* [Internet]. NIH Public Access; 2014 Sep 4 [cited 2017 Oct 21];55(5):708–22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25132174>
333. Ee L-S, McCannell KN, Tang Y, Fernandes N, Hardy WR, Green MR, et al. An Embryonic Stem Cell-Specific NuRD Complex Functions through Interaction with WDR5. *Stem cell reports* [Internet]. Elsevier; 2017 Jun 6 [cited 2017 Oct 21];8(6):1488–96. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28528697>
334. dos Santos RL, Tosti L, Radzisheuskaya A, Caballero IM, Kaji K, Hendrich B, et al. MBD3/NuRD facilitates induction of pluripotency in a context-dependent manner. *Cell Stem Cell* [Internet]. Elsevier; 2014 Jul 3 [cited 2017 Oct 21];15(1):102–10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24835571>
335. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* [Internet]. 2013 Oct 29 [cited 2015 Mar 4];110(44):17921–6. Available from: <http://www.pnas.org/content/110/44/17921.abstract>

336. Hoffman M (@michaelhoffman). This year for Halloween I'm dressing as a Super Enhancer. Part of quartet with Santa Claus, the Easter Bunny, and the Tooth Fairy. [Internet]. Twitter. 2017. Available from: <https://twitter.com/michaelhoffman/status/856536661433470976>
337. Robine N (@notSoJunkDNA). I fully expect these super-lncRNAs to be under the control of super-super-enhancers... (or should it be hyper-enhancer?) <https://twitter.com/segregatedxx/status/901054897872830468> [Internet]. Twitter. 2017. Available from: <https://twitter.com/notSoJunkDNA/status/901075042968240128>
338. Epigenetic Hulk. SUPERENHANCER NOMENCLATURE! HUH! WHAT IS IT GOOD FOR! ABSOLUTELY NOTHING! SAY IT AGAIN! <http://www.nature.com/ng/journal/v47/n1/full/ng.3167.html> [Internet]. Twitter. 2014. Available from: <https://twitter.com/epigenetichulk/status/550008601880694784>
339. Lynch VJ (@VinJLynch). A beautiful dissection of enhancer function yes, but evidence #SuperEnhancers are mech. distinct from enhancers? No. <http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3606.html> [Internet]. Twitter. 2016. Available from: <https://twitter.com/VinJLynch/status/750748932259602433>
340. Lowe D. Are Superenhancers Even Real? | In the Pipeline. In The Pipeline [Internet]. 2016 Sep [cited 2017 Oct 21]; Available from: <http://blogs.sciencemag.org/pipeline/archives/2016/09/13/are-superenhancers-even-real>
341. Pott S, Lieb JD. What are super-enhancers? Nat Genet [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Dec 30 [cited 2014 Dec 30];47(1):8–12. Available from: <http://dx.doi.org/10.1038/ng.3167>
342. Hamdan FH, Johnsen SA. Super enhancers – new analyses and perspectives on the low hanging fruit. Transcription [Internet]. Taylor & Francis; 2017 Oct 5 [cited 2017 Oct 21];00–00. Available from: <https://www.tandfonline.com/doi/full/10.1080/21541264.2017.1372044>
343. Hnisz D, Abraham B, Ihn Lee T, Young RA. Super-enhancers and methods of use thereof [Internet]. United States Patent and Trademark Office; WO2014066848 A1, 2013 [cited 2017 Oct 22]. Available from: <https://www.google.com/patents/WO2014066848A1?cl=en>
344. Shin HY, Willi M, Yoo KH, Zeng X, Wang C, Metser G, et al. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat Genet [Internet]. Nature Research; 2016 Jul 4 [cited 2017 Oct 14];48(8):904–11. Available from: <http://www.nature.com/doifinder/10.1038/ng.3606>
345. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol Cell [Internet]. Elsevier; 2017 Apr 20 [cited 2017 Oct 22];66(2):285–299.e5. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28416141>
346. Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, et al. DNA methylation dynamics of the human preimplantation embryo. Nature [Internet]. Nature Research; 2014 Jul 23 [cited 2017 Oct 11];511(7511):611–5. Available from: <http://www.nature.com/doifinder/10.1038/nature13581>

A APPENDICES

1.1 Tables

Table A.1 RT-qPCR primer sequences.

The nucleotide sequence of RT-qPCR primers used in this thesis. Target – gene targeted for amplification; Forward primer – the 5'-3' nucleotide sequence of the forward primer; Reverse primer – the 5'-3' nucleotide sequence of the reverse primer; Source – the lab in which a given primer pair was designed (VA - Dr Veronique Azuara, AJ – Dr Alice Jouneau).

Target	Forward primer	Reverse primer	Source
<i>Dnmt3b</i>	TTCAGTGACCAGTCCCTCAGACACGAA	TCAGAAGGCTGGAGACCTCCCTCTT	VA
<i>Esrrb</i>	ATGCGAGTACATGCTTAACGC	CATCCCCACTTTGAGGCATTT	AJ
<i>Fgf5</i>	TGTGTCTCAGGGGATTGTAGG	AGCTGTTTTCTTGGAATCTCTCC	AJ
<i>Klf13</i>	CCTGGCCTCAGACAAAGGG	ATTTCCCGTAAACTTTCTCGCA	AJ
<i>Klf2</i>	CGCCTCGGGTTCATTTC	AGCCTATCTTGCCGTCCTTT	VA
<i>Klf4</i>	GCAGTCACAAGTCCCTCTCTC	GACCTTCTTCCCTCTTTTGG	AJ
<i>Klf5</i>	GGTCCAGACAAGATGTGAAATGG	TTTATGCTCTGAAATTATCGGAACTG	VA
<i>Lefty1</i>	TCGATCAACCGCCAGTCC	CCGAACACTAGCAGGTGAGT	AJ
<i>Med13l</i>	GAACCCTGCCCTCTATGGAAA	TCTGACGACTCGGTGTATGAAG	AJ
<i>Nanog</i>	CTTTCACCTATTAAGGTGCTTGC	TGGCATCGGTTCATCATGGTAC	AJ
<i>Ncoa3</i>	TCTCAGCGCATCGACAACTT	GCCTGTAGTAAAAGCGGTCTTA	VA
<i>Otx2</i>	TATCTAAAGCAACCGCCTTACG	GCCCTAGTAAATGTCGTCCTCTC	AJ
<i>Pou5f1</i>	CAGCCAGACCACCATCTGTC	GTCTCCGATTGTCATATC	AJ
<i>Smarcad1</i>	GAAGGTAGCAGGGCAAACAC	TTTATTGCTGCGCGTTCTG	VA
<i>Sox2</i>	GCGGAGTGAAACTTTTGTCC	CGGGAAGCGTGTACTTATCCTT	AJ
<i>Tbx3</i>	GAACCTACCTGTTCCCGGAAA	AGTGTCTCGAAAACCCCTTTCG	AJ
<i>Tdh</i>	CTGCACAGAGCAAAGCCTCTA	TCAGGAGGTTAGCAAGTCCTAC	AJ
<i>Tet1</i>	GAGCCTGTTTCTCGATGTGG	CAAACCCACCTGAGGCTGTT	AJ
<i>Tet2</i>	AACCTGGCTACTGTCTATTGCTCCA	ATGTTCTGCTGGTCTCTGTGGGAA	AJ
<i>Tfcp2l1</i>	AGAGCATCTGCATTCATTACAGG	AAAAGGCACTCCCTTCTCACC	AJ
<i>Zfp42</i>	AGCTGGCCTAGAACTCACAG	GGCGATCCTGCTTCTTCTG	VA

Table A.2 GEO indexed sequencing and microarray data sets used in this study.

(S) = serum+LIF, (2i) = 2i+LIF.

Experiment	Target	Cell line	GEO #	Author
ATAC-seq		ESC (S)	GSE66581	Wu J, Huang B, Chen H, Yin Q et al. 2016[9]
Bisulphite-seq		EpiSC		Jouneau A
Bisulphite-seq		ESC (S)	GSE44092	Vincent JJ, Huang Y, Chen PY, Feng S et al. 2013[10]
ChIP-seq	CTCF	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	CTCF	ESC (S)	GSE18699	Goren A, Ozsolak F, Shores N, Ku M et al. 2010[16]
ChIP-seq	CTCF	ESC (S)	GSE30203	Stadler MB, Murr R, Burger L, Ivanek R et al. 2011[17]
ChIP-seq	CTCF	ESC (S)	GSE49847	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	ESRRB	ESC (S)	GSE28453	Adachi K, Nikaido I, Ohta H, Ohtsuka S et al. 2013[18]
ChIP-seq	ESRRB	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	H3K27ac	EpiLC	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	H3K27ac	EpiSC	GSE57407	Factor DC, Corradin O, Zentner GE, Saiakhova A et al. 2014[22]
ChIP-seq	H3K27ac	ESC (2i)	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	H3K27ac	ESC (S)	GSE24165	Creyghton MP, Cheng AW, Welstead GG, Kooistra T et al. 2010[23]
ChIP-seq	H3K27ac	ESC (S)	GSE43229	Das PP, Shao Z, Beyaz S, Apostolou E et al. 2014[24]
ChIP-seq	H3K27ac	ESC (S)	GSE39496	Ferrari KJ, Scelfo A, Jammula S, Cuomo A et al. 2014[25]
ChIP-seq	H3K27ac	ESC (S)	GSE53087	Kidder BL, Hu G, Zhao K. 2014[26]

ChIP-seq	H3K27ac	ESC (S)	GSE41903	Rowe HM, Kapopoulou A, Corsinotti A, Fasching L et al. 2013[27]
ChIP-seq	H3K27ac	ESC (S)	GSE29218	Shen Y, Yue F, McCleary DF, Ye Z et al. 2012[28]
ChIP-seq	H3K27ac	ESC (S)	GSE27841	Whyte WA, Bilodeau S, Orlando DA, Hoke HA et al. 2012[29]
ChIP-seq	H3K27ac	ESC (S)	GSE38596	Yu P, Xiao S, Xin X, Song CX et al. 2013[30]
ChIP-seq	H3K27ac	ESC (S)	GSE49847	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	H3K27ac	ESC (S)	GSE49847	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	H3K27me3	EpiSC	GSE57407	Factor DC, Corradin O, Zentner GE, Saiakhova A et al. 2014[22]
ChIP-seq	H3K27me3	ESC (S)	GSE43229	Das PP, Shao Z, Beyaz S, Apostolou E et al. 2014[24]
ChIP-seq	H3K27me3	ESC (S)	GSE23943	Marks H, Kalkan T, Menafr R, Denissov S et al. 2012[31]
ChIP-seq	H3K27me3	ESC (S)	GSE54955	Noordermeer D, Leleu M, Schorderet P, Joye E et al. 2014[32]
ChIP-seq	H3K27me3	ESC (S)	GSE25533	Vieux-Rochas M, Fabre PJ, Leleu M, Duboule D et al. 2015[33]
ChIP-seq	H3K27me3	ESC (S)	GSE38596	Tiwari VK, Stadler MB, Wirbelauer C, Paro R et al. 2011[34]
ChIP-seq	H3K27me3	ESC (S)	GSE49847	Tiwari VK, Burger L, Nikolettoulou V, Deogracias R et al. 2012[35]
ChIP-seq	H3K27me3	ESC (S)	GSE49847	Yu P, Xiao S, Xin X, Song CX et al. 2013[30]
ChIP-seq	H3K27me3	ESC (S)	GSE49847	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	H3K4me1	EpiLC	GSE56098	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	H3K4me1	EpiSC	GSE57407	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	H3K4me1	ESC (2i)	GSE56098	Factor DC, Corradin O, Zentner GE, Saiakhova A et al. 2014[22]
ChIP-seq	H3K4me1	ESC (S)	GSE24164	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	H3K4me1	ESC (S)	GSE11172	Creyghton MP, Cheng AW, Welstead GG, Kooistra T et al. 2010[23]
ChIP-seq	H3K4me1	ESC (S)	GSE38596	Meissner A, Mikkelsen TS, Gu H, Wernig M et al. 2008[37]
ChIP-seq	H3K4me1	ESC (S)	GSE49847	Yu P, Xiao S, Xin X, Song CX et al. 2013[30]
ChIP-seq	H3K4me3	ESC (S)	GSE15814	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	H3K4me3	ESC (S)	GSE12241	Marks H, Chow JC, Denissov S, Françoijis KJ et al. 2009[38]
ChIP-seq	H3K4me3	ESC (S)	GSE15519	Mikkelsen TS, Ku M, Jaffe DB, Issac B et al. 2007[36]
ChIP-seq	H3K4me3	ESC (S)	GSE49847	Goren A, Ozsolak F, Shores N, Ku M et al. 2010[16]
ChIP-seq	H3K4me3	ESC (S)	GSE49847	Rugg-Gunn PJ, Cox BJ, Ralston A, Rossant J. 2010[39]
ChIP-seq	KLF4	ESC (S)	GSE49848	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	KLF4	ESC (S)	GSE11431	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	KLF5	ESC (S)	GSE49848	Aksoy I, Giudice V, Delahaye E, Wianny F et al. 2014[44]
ChIP-seq	MED1	ESC (S)	GSE59696	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	MED1	ESC (S)	GSE22562	Aksoy I, Giudice V, Delahaye E, Wianny F et al. 2014[44]
ChIP-seq	MED1	ESC (S)	GSE49137	He Y, Schultz MD, Nery JR, Ecker JR
ChIP-seq	MED1	ESC (S)	GSE42474	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. 2010[46]
ChIP-seq	MED12	ESC (S)	GSE22562	Wang L, Du Y, Ward JM, Shimbo T et al. 2014[47]
ChIP-seq	NANOG	ESC (S)	GSE11431	Whyte WA, Orlando DA, Hnisz D, Abraham BJ et al. 2013[48]
ChIP-seq	NANOG	ESC (S)	GSE11724	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. 2010[46]
ChIP-seq	NANOG	ESC (S)	GSE44286	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	NANOG	ESC (S)	GSE49847	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
ChIP-seq	NCOA3	ESC (S)	GSE37262	Whyte WA, Orlando DA, Hnisz D, Abraham BJ et al. 2013[48]
ChIP-seq	NR5A2	ESC (S)	GSE19019	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	OCT4	EpiLC	GSE56098	Percharde M, Lavial F, Ng JH, Kumar V et al. 2012[50]
ChIP-seq	OCT4	ESC (2i)	GSE56098	Heng JC, Feng B, Han J, Jiang J et al. 2010[51]
ChIP-seq	OCT4	ESC (S)	GSE43275	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	OCT4	ESC (S)	GSE22934	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	OCT4	ESC (S)	GSE22934	Aksoy I, Jauch R, Chen J, Dyla M et al. 2013[52]
ChIP-seq	OCT4	ESC (S)	GSE11431	Ang YS, Tsai SY, Lee DF, Monk J et al. 2011[53]
ChIP-seq	OCT4	ESC (S)	GSE43229	Ang YS, Tsai SY, Lee DF, Monk J et al. 2011[53]
ChIP-seq	OCT4	ESC (S)	GSE34483	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	OCT4	ESC (S)	GSE36388	Das PP, Shao Z, Beyaz S, Apostolou E et al. 2014[24]
ChIP-seq	OCT4	ESC (S)	GSE11724	Hu G, Cui K, Northrup D, Liu C et al. 2013[54]
ChIP-seq	OCT4	ESC (S)	GSE11724	Jang H, Kim TW, Yoon S, Choi SY et al. 2012[55]
ChIP-seq	OCT4	ESC (S)	GSE49137	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
ChIP-seq	OCT4	ESC (S)	GSE44286	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
ChIP-seq	OCT4	ESC (S)	GSE49847	Wang L, Du Y, Ward JM, Shimbo T et al. 2014[47]
ChIP-seq	OTX2	EpiLC	GSE56098	Whyte WA, Orlando DA, Hnisz D, Abraham BJ et al. 2013[48]
ChIP-seq	OTX2	ESC (2i)	GSE56098	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	P300	EpiLC	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	P300	ESC (2i)	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	P300	ESC (S)	GSE49847	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
ChIP-seq	P300	ESC (S)	GSE11431	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	P300	ESC (S)	GSE24164	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	P300	ESC (S)	GSE22341	Creyghton MP, Cheng AW, Welstead GG, Kooistra T et al. 2010[23]
ChIP-seq	P300	ESC (S)	GSE22341	Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF et al. 2010[14]
ChIP-seq	P300	ESC (S)	GSE29218	Shen Y, Yue F, McCleary DF, Ye Z et al. 2012[28]
ChIP-seq	P300	ESC (S)	GSE41545	Song CX, Szulwach KE, Dai Q, Fu Y et al. 2013[56]
ChIP-seq	P300	ESC (S)	GSE49847	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
ChIP-seq	SMAD1	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	SMAD3	ESC (S)	GSE21614	Mullen AC, Orlando DA, Newman JJ, Lovén J et al. 2011[63]
ChIP-seq	SOX2	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]

ChIP-seq	SOX2	ESC (S)	GSE11724	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
ChIP-seq	SOX2	ESC (S)	GSE44286	Whyte WA, Orlando DA, Hnisz D, Abraham BJ et al. 2013[48]
ChIP-seq	STAT3	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	TCF7CP211	ESC (S)	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
ChIP-seq	TET1	ESC (S)	GSE26832	Wu H, D'Alessio AC, Ito S, Xia K et al. 2011[66]
ChIP-seq	TET1	ESC (S)	GSE26832	Wu H, D'Alessio AC, Ito S, Wang Z et al. 2011[67]
				Wu H, D'Alessio AC, Ito S, Xia K et al. 2011[66]
				Wu H, D'Alessio AC, Ito S, Wang Z et al. 2011[67]
				Xu Y, Wu F, Tan L, Kong L et al. 2011[68]
ChIP-seq	TET1	ESC (S)	GSE28532	Chen Q, Chen Y, Bian C, Fujiki R et al. 2013[69]
ChIP-seq	TET2	ESC (S)	GSE41720	Song CX, Szulwach KE, Dai Q, Fu Y et al. 2013[56]
DIP-seq	5FC	ESC (S)	GSE41545	Song CX, Szulwach KE, Dai Q, Fu Y et al. 2013[56]
DIP-seq	5HMC	ESC (S)	GSE41545	Song CX, Szulwach KE, Dai Q, Fu Y et al. 2013[56]
DIP-seq	5HMC	ESC (S)	GSE28532	Xu Y, Wu F, Tan L, Kong L et al. 2011[68]
DIP-seq	5HMC	ESC (S)	GSE38596	Yu P, Xiao S, Xin X, Song CX et al. 2013[30]
DNase-seq		EpiSC	GSE57406	Factor DC, Corradin O, Zentner GE, Saiakhova A et al. 2014[22]
FAIRE-seq		EpiLC	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
FAIRE-seq		EpiSC	GSE58520	Murtha M, Dailey L
FAIRE-seq		ESC (2i)	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
FAIRE-seq		ESC (S)	GSE58520	Murtha M, Dailey L
MeDIP-seq		ESC (S)	GSE31343	Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG. 2011[72]
MethylCap		ESC (S)	GSE31343	Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG. 2011[72]
Microarray		EpiSC	GSE46227	Kojima Y, Kaufman-Francis K, Studdert JB, Steiner KA et al. 2014[73]
Microarray		EpiSC	GSE46227	Kojima Y, Kaufman-Francis K, Studdert JB, Steiner KA et al. 2014[73]
Microarray	Esrrb KO	ESC (S)		Niwa H
Microarray	Esrrb RNAi	ESC (S)	GSE26520	Nishiyama A, Sharov AA, Piao Y, Amano M et al. 2013[74]
Microarray	Esrrb RNAi	ESC (S)	GSE13212	Feng B, Jiang J, Kraus P, Ng JH et al. 2009[75]
Microarray	Esrrb RNAi	ESC (S)	GSE34170	Hutchins AP, Choo SH, Mistri TK, Rahmani M et al. 2013[76]
Microarray	Esrrb RNAi	ESC (S)	GSE4679	Ivanova N, Dobrin R, Lu R, Kotenko I et al. 2006[77]
Microarray	Nanog RNAi	ESC (S)	GSE4679	Ivanova N, Dobrin R, Lu R, Kotenko I et al. 2006[77]
Microarray	Nanog RNAi	ESC (S)	GSE4189	Loh YH, Wu Q, Chew JL, Vega VB et al. 2006[78]
Microarray	Nanog RNAi	ESC (S)	GSE53529	Stuart HT, van Oosten AL, Radzisheuskaya A, Martello G et al. 2014[79]
Microarray	Nanog RNAi	ESC (S)	GSE14559	Nishiyama A, Xin L, Sharov AA, Thomas M et al. 2009[80]
Microarray	Nanog RNAi	ESC (S)	GSE26520	Nishiyama A, Sharov AA, Piao Y, Amano M et al. 2013[74]
Microarray	Ncoa3 RNAi	ESC (S)	GSE37262	Percharde M, Lavial F, Ng JH, Kumar V et al. 2012[50]
Microarray	Ncoa3 RNAi	ESC (S)	GSE26520	Nishiyama A, Sharov AA, Piao Y, Amano M et al. 2013[74]
Microarray	Ncoa3 RNAi	ESC (S)	GSE50017	Chitilian JM, Thillainadesan G, Manias JL, Chang WY et al. 2014[81]
Microarray	Pou5f1 RNAi	ESC (S)	GSE44287	Whyte WA, Orlando DA, Hnisz D, Abraham BJ et al. 2013[48]
Microarray	shRNA screen	ESC (S)	GSE26520	Nishiyama A, Sharov AA, Piao Y, Amano M et al. 2013[74]
RNA-seq		Embryo		Bernardo et al (in submission)
RNA-seq		EpiLC	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
RNA-seq		EpiSC	GSE47793	Veillard AC, Marks H, Bernardo AS, Jouneau L et al. 2014[82]
RNA-seq		ESC (2i)	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
RNA-seq		ESC (2i)	GSE23943	Marks H, Kalkan T, Menafrá R, Denissov S et al. 2012[31]
RNA-seq		ESC (S)	GSE56098	Buecker C, Srinivasan R, Wu Z, Calo E et al. 2014[21]
RNA-seq		ESC (S)	GSE23943	Marks H, Kalkan T, Menafrá R, Denissov S et al. 2012[31]

Table A.3 The ChIP-seq data visualised in figure 3.2A.

IP target	GEO #	Author
ESRRB	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
H3K27ac	GSE31039	Yue F, Cheng Y, Breschi A, Vierstra J et al. 2014[12]
H3K4me1	GSE24164	Creyghton MP, Cheng AW, Welstead GG, Kooistra T et al. 2010[23]
KLF4	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]
LRH1	GSE19019	Heng JC, Feng B, Han J, Jiang J et al. 2010[51]
MED1	GSE22562	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. 2010[46]
MED12	GSE22562	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. 2010[46]
NANOG	GSE11724	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
NCOA3	GSE37262	Percharde M, Lavial F, Ng JH, Kumar V et al. 2012[50]
OCT4	GSE22934	Ang YS, Tsai SY, Lee DF, Monk J et al. 2011[53]
P300	GSE22341	Schnetzer MP, Handoko L, Akhtar-Zaidi B, Bartels CF et al. 2010[14]
SMAD3	GSE21614	Mullen AC, Orlando DA, Newman JJ, Lovén J et al. 2011[63]
SOX2	GSE11724	Marson A, Levine SS, Cole MF, Frampton GM et al. 2008[36]
STAT3	GSE11431	Chen X, Xu H, Yuan P, Fang F et al. 2008[15]

Table A.4 ESC Super Enhancers and their associated genes, by Super Enhancelet complementation and class.

Class I = Maintained, Class II = Silenced.

Chromosome	Start	End	Associated gene	PU/DM/Both	Class
1	34130107	34134640	<i>Dst</i>	Both	I
1	36070190	36074608	<i>Hs6st1</i>	Both	I
1	36111164	36118698	<i>Hs6st1</i>	Both	I
1	37039139	37045411	<i>Tmem131</i>	Both	I
1	72839563	72858199	<i>Igfbp2</i>	Both	I
1	121201424	121202481	<i>Inhbb</i>	Both	I
1	137071028	137096284	<i>Gpr37l1</i>	Both	I
1	138586629	138593131	<i>Zfp281</i>	Both	I
1	138841643	138850970	<i>Nr5a2</i>	PU	I
1	169201106	169220423	<i>Uck2</i>	Both	I
1	182854521	182864307	<i>Lefty1</i>	PU	I
1	183948212	183961841	<i>Enah</i>	Both	I
2	32008891	32030736	<i>Prrc2b</i>	Both	I
2	33282029	33300860	<i>Zbtb34</i>	Both	I
2	71488013	71494617	<i>Gm1631</i>	PU	I
2	71575856	71583914	<i>Gm1631</i>	PU	I
2	152552277	152563676	<i>Id1</i>	Both	I
2	154242651	154254374	<i>Cbfa2t2</i>	Both	I
2	162877048	162893236	<i>Mybl2</i>	Both	I
2	165981373	165983444	<i>Sulf2</i>	Both	I
2	168589688	168617170	<i>Sall4</i>	Both	I
3	34544904	34553511	<i>Sox2</i>	Both	I
3	34633687	34660705	<i>Sox2</i>	PU	I
3	95455034	95468269	<i>Mcl1</i>	Both	I
3	96479158	96484864	<i>Ankrd35</i>	Both	I
3	103008304	103019058	<i>Bcas2</i>	Both	I
4	118743867	118745786	<i>Slc2a1</i>	Both	I
4	123300547	123303179	<i>Macf1</i>	Both	I
4	130178808	130180168	<i>Pum1</i>	PU	I
4	137148873	137153839	<i>Usp48</i>	Both	I
4	137329436	137357766	<i>Alpl</i>	Both	I
4	138000554	138006368	<i>Camk2n1</i>	Both	I
4	140826072	140840922	<i>Epha2</i>	Both	I
4	141721916	141726166	<i>Kazn</i>	Both	I
4	147459254	147463850	<i>Agtrap</i>	Both	I
4	154563584	154564383	<i>Ski</i>	PU	I
5	33873714	33880481	<i>Fam53a</i>	Both	I
5	116845764	116860853	<i>Hspb8</i>	Both	I
5	118884660	118896412	<i>Med13l</i>	Both	I
5	123584659	123590728	<i>Rhof</i>	PU	I
6	39370384	39371286	<i>Mkrm1</i>	PU	I
6	39395571	39396779	<i>Mkrm1</i>	PU	I
6	64961359	64985161	<i>Smardc1</i>	Both	I
6	122290093	122293017	<i>Phc1</i>	PU	I
6	122640118	122657871	<i>Nanog</i>	Both	I
6	122714316	122720862	<i>Slc2a3</i>	Both	I
6	143047309	143065758	<i>5730419I09Rik</i>	Both	I
6	145223385	145225674	<i>Kras</i>	Both	I
7	3193004	3218183	<i>Mir290</i>	Both	I
7	52806853	52814768	<i>Hsd17b14</i>	Both	I
7	56592909	56604632	<i>Nav2</i>	Both	I
7	71092246	71102481	<i>Klf13</i>	Both	I
7	87159908	87169963	<i>Zfp710</i>	Both	I
7	87333420	87345334	<i>Sema4b</i>	Both	I
7	91027196	91051830	<i>Mesdc2</i>	Both	I
7	140304156	140307245	<i>Ctbp2</i>	Both	I
7	147131117	147136231	<i>Uff1</i>	Both	I
7	152036872	152050716	<i>Fgf4</i>	Both	I
8	12499468	12504771	<i>Spaca7</i>	PU	I
8	35023426	35027483	<i>Rbpms</i>	Both	I
8	37602064	37613850	<i>6430573F11Rik</i>	PU	I
8	91514813	91540176	<i>Sall1</i>	Both	I

8	93351924	93355292	<i>Chd9</i>	PU	I
9	58119837	58128504	<i>Stoml1</i>	Both	I
9	110849422	110863371	<i>Lrrc2</i>	Both	I
9	114458126	114474355	<i>Trim71</i>	Both	I
9	120585871	120600072	<i>Rpl14</i>	Both	I
10	20802131	20830236	<i>Ahi1</i>	Both	I
10	39977900	39978752	<i>Gtf3c6</i>	PU	I
10	59420365	59437537	<i>Ddit4</i>	Both	I
10	62346394	62361563	<i>Tet1</i>	Both	I
10	66380351	66383761	<i>Reep3</i>	Both	I
10	66546199	66564235	<i>Reep3</i>	Both	I
10	75335464	75345568	<i>Derl3</i>	Both	I
10	79508474	79515168	<i>Gpx4</i>	Both	I
10	85002060	85006553	<i>Btbd11</i>	Both	I
11	9015537	9017663	<i>Upp1</i>	PU	I
11	12357626	12370205	<i>Cobl</i>	Both	I
11	52173182	52184686	<i>Vdac1</i>	Both	I
11	54767341	54785832	<i>Tnfr1</i>	Both	I
11	66824791	66838230	<i>Tmem220</i>	Both	I
11	77697704	77718786	<i>Pipox</i>	Both	I
11	88481360	88491812	<i>Msi2</i>	Both	I
11	97517673	97524159	<i>Mlt6</i>	Both	I
11	98823511	98826466	<i>Rara</i>	Both	I
11	104150171	104167544	<i>Mapt</i>	Both	I
11	116943025	116953583	<i>2810008D09Rik</i>	Both	I
11	117833701	117838253	<i>Socs3</i>	Both	I
12	12790432	12795881	<i>Mycn</i>	Both	I
12	12933791	12950936	<i>Mycn</i>	Both	I
12	55407498	55415046	<i>Gm7550</i>	PU	I
12	56587347	56607146	<i>Nfkb1a</i>	Both	I
12	57385208	57400114	<i>Mbip</i>	Both	I
12	103940487	103953004	<i>Itpk1</i>	Both	I
13	64069823	64082322	<i>Hsd17b3</i>	PU	I
13	96295094	96306119	<i>F2r1</i>	Both	I
13	98052562	98062842	<i>Enc1</i>	Both	I
13	98202400	98225162	<i>Enc1</i>	Both	I
13	110418702	110442750	<i>Mir1904</i>	Both	I
14	49273113	49283200	<i>Otx2</i>	Both	I
14	71022659	71035930	<i>Epb4.9</i>	Both	I
14	76894682	76915946	<i>Tsc22d1</i>	Both	I
14	106250319	106260753	<i>Spry2</i>	Both	I
14	106296486	106304433	<i>Spry2</i>	Both	I
15	25654102	25704265	<i>Fam134b</i>	Both	I
15	61918415	61924748	<i>Pvt1</i>	Both	I
15	77168852	77187251	<i>Rbfox2</i>	Both	I
15	97198605	97227633	<i>Amigo2</i>	Both	I
16	8758173	8779472	<i>Usp7</i>	Both	I
16	23099373	23103471	<i>Eif4a2</i>	PU	I
16	29657509	29668114	<i>Opa1</i>	Both	I
16	84769173	84780686	<i>Jam2</i>	Both	I
17	26631721	26648689	<i>Dusp1</i>	Both	I
17	29209618	29218426	<i>Gm16197</i>	Both	I
17	29587776	29588942	<i>Pim1</i>	Both	I
17	35639211	35642435	<i>Pou5f1</i>	Both	I
17	37209046	37217726	<i>Olfr90</i>	Both	I
17	47640414	47649043	<i>Ccnd3</i>	Both	I
17	66818723	66836409	<i>1110012J17Rik</i>	Both	I
17	71096763	71100905	<i>Tgif1</i>	Both	I
18	38538325	38551037	<i>Ndfip1</i>	Both	I
18	38760823	38761958	<i>Spry4</i>	PU	I
18	38788269	38796942	<i>Spry4</i>	Both	I
18	75520332	75527277	<i>Smad7</i>	Both	I
19	5835881	5847014	<i>Neat1</i>	Both	I
19	23207455	23208806	<i>Klf9</i>	PU	I
19	53523440	53535319	<i>5830416P10Rik</i>	PU	I
X	7578969	7597907	<i>Glod5</i>	Both	I
X	50098631	50114110	<i>Kis2</i>	PU	I
1	13049615	13094765	<i>Prdm14</i>	DM	II
1	72260528	72261272	<i>Mreg</i>	DM	II
1	84857219	84887132	<i>Fbxo36</i>	DM	II
1	91766947	91773527	<i>Gbx2</i>	DM	II
1	120538712	120545414	<i>Tfcp2l1</i>	DM	II

1	120971968	120973737	<i>Gli2</i>	DM	II
1	121295085	121296031	<i>Inhbb</i>	DM	II
1	154939892	154943709	<i>Lamc2</i>	DM	II
1	168054897	168073079	<i>Gpa33</i>	DM	II
2	20574602	20591747	<i>Etl4</i>	DM	II
2	30913257	30925299	<i>Fnbp1</i>	DM	II
2	152002668	152003777	<i>Tcf15</i>	DM	II
2	162856904	162860933	<i>Ift52</i>	DM	II
3	9641461	9655131	<i>Zfp704</i>	DM	II
3	88375442	88380083	<i>Ssr2</i>	DM	II
3	96380383	96382115	<i>Polr3gl</i>	DM	II
3	129247012	129261362	<i>Elovl6</i>	DM	II
3	133181431	133197648	<i>Tet2</i>	DM	II
3	135208956	135210744	<i>Manba</i>	DM	II
4	55469259	55491081	<i>Klf4</i>	DM	II
4	125211671	125223450	<i>Ftl2</i>	DM	II
4	126875757	126879027	<i>Dlgap3</i>	DM	II
4	130195646	130196547	<i>Pum1</i>	DM	II
4	141120768	141126477	<i>B330016D10Rik</i>	DM	II
4	141616653	141627603	<i>Gm10565</i>	DM	II
4	154537213	154538078	<i>Ski</i>	DM	II
5	53933177	53947327	<i>Rbpj</i>	DM	II
5	65255735	65256794	<i>Klf3</i>	DM	II
5	118951444	118960269	<i>Med13l</i>	DM	II
5	120029649	120037063	<i>Tbx3</i>	DM	II
5	120129592	120171482	<i>Tbx3</i>	DM	II
5	135417523	135421698	<i>Cldn4</i>	DM	II
6	31834643	31852445	<i>1700012A03Rik</i>	DM	II
6	67061148	67064202	<i>Gadd45a</i>	DM	II
6	91640161	91661247	<i>Slc6a6</i>	DM	II
6	122612514	122614260	<i>Dppa3</i>	DM	II
6	125383335	125398024	<i>Cd9</i>	DM	II
6	142458188	142461905	<i>Ldhb</i>	DM	II
7	13599334	13600325	<i>Zbtb45</i>	DM	II
7	30982397	30983339	<i>Capns1</i>	DM	II
7	31248315	31250619	<i>Nphs1as</i>	DM	II
7	38812914	38816123	<i>C80913</i>	DM	II
7	86355826	86368339	<i>Abhd2</i>	DM	II
7	87274999	87276022	<i>Idh2</i>	DM	II
7	119831735	119835688	<i>Tead1</i>	DM	II
8	37642521	37671979	<i>Dlc1</i>	DM	II
8	44405736	44406755	<i>Zfp42</i>	DM	II
8	74834685	74840663	<i>Klf2</i>	DM	II
8	87996475	87997654	<i>Gpt2</i>	DM	II
9	56382386	56395769	<i>Hmg20a</i>	DM	II
9	78207143	78223442	<i>Dppa5a</i>	DM	II
9	121244501	121254102	<i>Trak1</i>	DM	II
10	21546502	21549691	<i>Sgk1</i>	DM	II
10	21700576	21708946	<i>Sgk1</i>	DM	II
10	44110139	44112766	<i>Mir1929</i>	DM	II
10	75400370	75401358	<i>Chchd10</i>	DM	II
10	76655655	76662360	<i>Col18a1</i>	DM	II
11	8466451	8486876	<i>Tns3</i>	DM	II
11	33427175	33451476	<i>Ranbp17</i>	DM	II
11	62324296	62327251	<i>Cenpv</i>	DM	II
11	66733372	66746990	<i>Pirt</i>	DM	II
11	69517060	69522803	<i>BC096441</i>	DM	II
11	102190649	102193692	<i>Ubtg</i>	DM	II
11	107296669	107310982	<i>Pitpnc1</i>	DM	II
12	12810177	12811020	<i>Mycn</i>	DM	II
12	87807046	87820319	<i>Esrrb</i>	DM	II
12	87839385	87846192	<i>Esrrb</i>	DM	II
12	88239069	88245155	<i>6430527G18Rik</i>	DM	II
12	111655417	111656705	<i>Ppp2r5c</i>	DM	II
12	111709296	111710794	<i>Ppp2r5c</i>	DM	II
12	111725920	111743677	<i>Ppp2r5c</i>	DM	II
14	22293688	22308989	<i>Myst4</i>	DM	II
14	55704349	55705463	<i>Zfhx2</i>	DM	II
14	64118817	64131901	<i>Tdh</i>	DM	II
14	65251303	65269514	<i>Kif13b</i>	DM	II
14	77015215	77030315	<i>1700108F19Rik</i>	DM	II
14	99738540	99755307	<i>Klf5</i>	DM	II

15	88539016	88539831	<i>Brd1</i>	DM	II
15	97422878	97425328	<i>Rpap3</i>	DM	II
15	103349226	103353500	<i>Ppp1r1a</i>	DM	II
17	10549089	10570838	<i>Qk</i>	DM	II
17	31939569	31956756	<i>Sik1</i>	DM	II
17	37110202	37134996	<i>H2-M5</i>	DM	II
17	45593477	45596503	<i>Gm9104</i>	DM	II
17	71177302	71179956	<i>Tgif1</i>	DM	II
17	71213804	71222433	<i>Tgif1</i>	DM	II
17	71241991	71250610	<i>Tgif1</i>	DM	II
18	36412873	36414154	<i>Pura</i>	DM	II
18	61787544	61788400	<i>Mir145</i>	DM	II
18	75504155	75505202	<i>Smad7</i>	DM	II
18	75738693	75745073	<i>Ctif</i>	DM	II
19	21858770	21866770	<i>Tmem2</i>	DM	II
19	23139991	23170189	<i>Klf9</i>	DM	II
19	25553498	25564092	<i>Dmrt1</i>	DM	II
1	182818684	182819554	<i>Lefty2</i>	NA	NA
4	98507649	98514709	<i>Kank4</i>	NA	NA
5	113758941	113775389	<i>Mir469</i>	NA	NA
6	83839914	83844315	<i>Zfml</i>	NA	NA
7	4772296	4777612	<i>Ube2s</i>	NA	NA
8	87174072	87174643	<i>Ier2</i>	NA	NA
18	35202713	35203454	NA	NA	NA
18	40467587	40468140	<i>Kctd16</i>	NA	NA

Table A.5 Gene Ontology terms enriched in the Class I (Maintained) ESC Super Enhancer associated gene set.

Gene Ontology (GO) was performed using the PANTHERDB over-representation test. P-values were adjusted for multiple hypothesis testing using the Bonferroni method.

Parent Term	GO biological process complete	Mus musculus - REFLIST (22221)	Class I (108)	Class I (expected)	Class I (over/under)	Class I (fold Enrichment)	Class I (P-value)	Class I (-log ₁₀ (P-value))
cellular response to leukemia inhibitory factor (GO:1990830)	cellular response to leukemia inhibitory factor (GO:1990830)	127	16	0.62	+	25.92	3.41E-14	13.47
cellular response to leukemia inhibitory factor (GO:1990830)	response to leukemia inhibitory factor (GO:1990823)	127	16	0.62	+	25.92	3.41E-14	13.47
cellular response to leukemia inhibitory factor (GO:1990830)	cellular response to organic substance (GO:0071310)	1571	31	7.64	+	4.06	8.72E-08	7.06
cellular response to leukemia inhibitory factor (GO:1990830)	response to organic substance (GO:0010033)	2255	37	10.96	+	3.38	1.05E-07	6.98
cellular response to leukemia inhibitory factor (GO:1990830)	cellular response to chemical stimulus (GO:0070887)	1977	33	9.61	+	3.43	1.29E-06	5.89
cellular response to leukemia inhibitory factor (GO:1990830)	cellular response to cytokine stimulus (GO:0071345)	549	18	2.67	+	6.75	1.66E-06	5.78
cellular response to leukemia inhibitory factor (GO:1990830)	response to cytokine (GO:0034097)	648	19	3.15	+	6.03	3.14E-06	5.50
cellular response to leukemia inhibitory factor (GO:1990830)	negative regulation of response to stimulus (GO:0048585)	1383	24	6.72	+	3.57	3.62E-04	3.44
cellular response to leukemia inhibitory factor (GO:1990830)	response to chemical (GO:0042221)	3274	38	15.91	+	2.39	9.16E-04	3.04
cellular response to leukemia inhibitory factor (GO:1990830)	regulation of response to stimulus (GO:0048583)	3338	38	16.22	+	2.34	1.52E-03	2.82
cellular response to leukemia inhibitory factor (GO:1990830)	cellular response to stimulus (GO:0051716)	6038	52	29.35	+	1.77	2.33E-02	1.63
endodermal cell fate specification (GO:0001714)	embryonic morphogenesis (GO:0048598)	598	23	2.91	+	7.91	1.30E-10	9.89
endodermal cell fate specification (GO:0001714)	embryo development (GO:0009790)	1056	27	5.13	+	5.26	7.61E-09	8.12
endodermal cell fate specification (GO:0001714)	tissue development (GO:0009888)	1522	31	7.4	+	4.19	3.89E-08	7.41
endodermal cell fate specification (GO:0001714)	developmental process (GO:0032502)	5122	57	24.89	+	2.29	1.79E-07	6.75
endodermal cell fate specification (GO:0001714)	cell differentiation (GO:0030154)	3427	44	16.66	+	2.64	1.88E-06	5.73

endodermal cell fate specification (GO:0001714)	anatomical structure development (GO:0048856)	4795	53	23.3	+	2.27	2.26E-06	5.65
endodermal cell fate specification (GO:0001714)	anatomical structure morphogenesis (GO:0009653)	2031	33	9.87	+	3.34	2.59E-06	5.59
endodermal cell fate specification (GO:0001714)	multicellular organism development (GO:0007275)	4506	51	21.9	+	2.33	2.76E-06	5.56
endodermal cell fate specification (GO:0001714)	cellular developmental process (GO:0048869)	3496	44	16.99	+	2.59	3.60E-06	5.44
endodermal cell fate specification (GO:0001714)	anatomical structure formation involved in morphogenesis (GO:0048646)	851	19	4.14	+	4.59	2.52E-04	3.60
endodermal cell fate specification (GO:0001714)	cellular process (GO:0009987)	13735	92	66.76	+	1.38	7.27E-04	3.14
endodermal cell fate specification (GO:0001714)	gastrulation (GO:0007369)	144	8	0.7	+	11.43	5.22E-03	2.28
endodermal cell fate specification (GO:0001714)	endodermal cell fate specification (GO:0001714)	6	3	0.03	+	> 100	3.34E-02	1.48
endodermal cell fate specification (GO:0001714)	multicellular organismal process (GO:0032501)	7032	57	34.18	+	1.67	3.66E-02	1.44
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	negative regulation of macromolecule metabolic process (GO:0010605)	2276	34	11.06	+	3.07	1.15E-05	4.94
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	negative regulation of macromolecule biosynthetic process (GO:0010558)	1334	22	6.48	+	3.39	3.50E-03	2.46
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	negative regulation of cellular biosynthetic process (GO:0031327)	1405	22	6.83	+	3.22	8.32E-03	2.08
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	negative regulation of biosynthetic process (GO:0009890)	1435	22	6.97	+	3.15	1.18E-02	1.93
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	1283	20	6.24	+	3.21	3.01E-02	1.52
negative regulation of intracellular signal transduction (GO:1902532)	negative regulation of biological process (GO:0048519)	4650	54	22.6	+	2.39	1.84E-07	6.74
negative regulation of intracellular signal transduction (GO:1902532)	regulation of signal transduction (GO:0009966)	2482	36	12.06	+	2.98	6.90E-06	5.16
negative regulation of intracellular signal transduction (GO:1902532)	negative regulation of signal transduction (GO:0009968)	1067	22	5.19	+	4.24	7.21E-05	4.14
negative regulation of intracellular signal transduction (GO:1902532)	negative regulation of cell communication (GO:0010648)	1193	23	5.8	+	3.97	1.08E-04	3.97
negative regulation of intracellular signal transduction (GO:1902532)	negative regulation of signaling (GO:0023057)	1197	23	5.82	+	3.95	1.15E-04	3.94
negative regulation of intracellular signal transduction (GO:1902532)	regulation of cell communication (GO:0010646)	2810	36	13.66	+	2.64	1.89E-04	3.72
negative regulation of intracellular signal transduction (GO:1902532)	regulation of signaling (GO:0023051)	2832	36	13.76	+	2.62	2.32E-04	3.63
negative regulation of intracellular signal transduction (GO:1902532)	regulation of intracellular signal transduction (GO:1902531)	1555	23	7.56	+	3.04	1.19E-02	1.92
negative regulation of intracellular signal transduction (GO:1902532)	negative regulation of intracellular signal transduction (GO:1902532)	494	12	2.4	+	5	4.77E-02	1.32
negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	negative regulation of cellular process (GO:0048523)	4219	52	20.51	+	2.54	5.92E-08	7.23
negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	negative regulation of metabolic process (GO:0009892)	2522	36	12.26	+	2.94	1.07E-05	4.97

negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	negative regulation of nitrogen compound metabolic process (GO:0051172)	2154	32	10.47	+	3.06	4.77E-05	4.32
negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	1344	21	6.53	+	3.21	1.59E-02	1.80
neural tube closure (GO:0001843)	tissue morphogenesis (GO:0048729)	583	20	2.83	+	7.06	6.45E-08	7.19
neural tube closure (GO:0001843)	tube development (GO:0035295)	627	20	3.05	+	6.56	2.34E-07	6.63
neural tube closure (GO:0001843)	epithelium development (GO:0060429)	987	24	4.8	+	5	4.95E-07	6.31
neural tube closure (GO:0001843)	epithelial tube morphogenesis (GO:0060562)	347	15	1.69	+	8.89	1.51E-06	5.82
neural tube closure (GO:0001843)	tube morphogenesis (GO:0035239)	388	15	1.89	+	7.95	6.88E-06	5.16
neural tube closure (GO:0001843)	morphogenesis of an epithelium (GO:0002009)	481	16	2.34	+	6.84	1.52E-05	4.82
neural tube closure (GO:0001843)	neural tube development (GO:0021915)	184	11	0.89	+	12.3	1.76E-05	4.75
neural tube closure (GO:0001843)	neural tube closure (GO:0001843)	106	9	0.52	+	17.47	2.77E-05	4.56
neural tube closure (GO:0001843)	tube closure (GO:0060606)	107	9	0.52	+	17.31	3.01E-05	4.52
neural tube closure (GO:0001843)	primary neural tube formation (GO:0014020)	113	9	0.55	+	16.39	4.80E-05	4.32
neural tube closure (GO:0001843)	chordate embryonic development (GO:0043009)	698	18	3.39	+	5.31	6.98E-05	4.16
neural tube closure (GO:0001843)	embryo development ending in birth or egg hatching (GO:0009792)	707	18	3.44	+	5.24	8.49E-05	4.07
neural tube closure (GO:0001843)	neural tube formation (GO:0001841)	127	9	0.62	+	14.58	1.30E-04	3.89
neural tube closure (GO:0001843)	embryonic epithelial tube formation (GO:0001838)	148	9	0.72	+	12.51	4.73E-04	3.33
neural tube closure (GO:0001843)	epithelial tube formation (GO:0072175)	154	9	0.75	+	12.02	6.60E-04	3.18
neural tube closure (GO:0001843)	tube formation (GO:0035148)	168	9	0.82	+	11.02	1.37E-03	2.86
neural tube closure (GO:0001843)	morphogenesis of embryonic epithelium (GO:0016331)	172	9	0.84	+	10.77	1.66E-03	2.78
neural tube closure (GO:0001843)	system development (GO:0048731)	3943	41	19.16	+	2.14	4.69E-03	2.33
neural tube closure (GO:0001843)	nervous system development (GO:0007399)	2093	28	10.17	+	2.75	4.91E-03	2.31
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	negative regulation of cellular metabolic process (GO:0031324)	2316	36	11.26	+	3.2	1.02E-06	5.99
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of nucleobase-containing compound metabolic process (GO:0019219)	3634	45	17.66	+	2.55	3.44E-06	5.46
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of transcription from RNA polymerase II promoter (GO:0006357)	1837	31	8.93	+	3.47	4.27E-06	5.37
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of macromolecule biosynthetic process (GO:0010556)	3580	44	17.4	+	2.53	7.75E-06	5.11
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of cellular process (GO:0048522)	4836	52	23.5	+	2.21	1.08E-05	4.97
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of RNA metabolic process (GO:0051254)	1477	27	7.18	+	3.76	1.38E-05	4.86

positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of gene expression (GO:0010468)	3837	45	18.65	+	2.41	2.03E-05	4.69
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of RNA metabolic process (GO:0051252)	3273	41	15.91	+	2.58	2.30E-05	4.64
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	1117	23	5.43	+	4.24	3.19E-05	4.50
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of cellular biosynthetic process (GO:0031326)	3770	44	18.32	+	2.4	4.01E-05	4.40
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of cellular macromolecule biosynthetic process (GO:2000112)	3478	42	16.9	+	2.48	4.01E-05	4.40
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of biosynthetic process (GO:0009889)	3826	44	18.6	+	2.37	6.36E-05	4.20
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of nucleic acid-templated transcription (GO:1903506)	3109	39	15.11	+	2.58	6.61E-05	4.18
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of RNA biosynthetic process (GO:2001141)	3115	39	15.14	+	2.58	6.98E-05	4.16
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of macromolecule metabolic process (GO:0010604)	2862	37	13.91	+	2.66	8.62E-05	4.06
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of nucleic acid-templated transcription (GO:1903508)	1404	25	6.82	+	3.66	1.07E-04	3.97
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of transcription, DNA-templated (GO:0045893)	1404	25	6.82	+	3.66	1.07E-04	3.97
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of RNA biosynthetic process (GO:1902680)	1405	25	6.83	+	3.66	1.09E-04	3.96
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of biological process (GO:0048518)	5342	53	25.96	+	2.04	1.27E-04	3.90
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of cellular metabolic process (GO:0031325)	2909	37	14.14	+	2.62	1.33E-04	3.88
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of gene expression (GO:0010628)	1760	28	8.55	+	3.27	1.34E-04	3.87
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of nitrogen compound metabolic process (GO:0051173)	2788	36	13.55	+	2.66	1.54E-04	3.81
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	regulation of transcription, DNA-templated (GO:0006355)	3094	38	15.04	+	2.53	2.02E-04	3.69
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of metabolic process (GO:0009893)	3106	38	15.1	+	2.52	2.24E-04	3.65
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of nucleobase-containing compound metabolic process (GO:0045935)	1727	27	8.39	+	3.22	3.70E-04	3.43
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of macromolecule biosynthetic process (GO:0010557)	1646	25	8	+	3.12	2.28E-03	2.64
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of cellular biosynthetic process (GO:0031328)	1780	25	8.65	+	2.89	9.64E-03	2.02
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	positive regulation of biosynthetic process (GO:0009891)	1814	25	8.82	+	2.84	1.36E-02	1.87
regulation of cell differentiation (GO:0045595)	regulation of cell differentiation (GO:0045595)	1673	24	8.13	+	2.95	1.15E-02	1.94

regulation of cell differentiation (GO:0045595)	regulation of developmental process (GO:0050793)	2357	29	11.46	+	2.53	1.61E-02	1.79
regulation of MAPK cascade (GO:0043408)	regulation of cellular metabolic process (GO:0031323)	5423	56	26.36	+	2.12	6.63E-06	5.18
regulation of MAPK cascade (GO:0043408)	regulation of macromolecule metabolic process (GO:0060255)	5312	55	25.82	+	2.13	9.90E-06	5.00
regulation of MAPK cascade (GO:0043408)	regulation of nitrogen compound metabolic process (GO:0051171)	5169	54	25.12	+	2.15	1.16E-05	4.94
regulation of MAPK cascade (GO:0043408)	regulation of metabolic process (GO:0019222)	5756	57	27.98	+	2.04	2.14E-05	4.67
regulation of MAPK cascade (GO:0043408)	regulation of primary metabolic process (GO:0080090)	5294	54	25.73	+	2.1	2.87E-05	4.54
regulation of MAPK cascade (GO:0043408)	regulation of cellular protein metabolic process (GO:0032268)	2289	29	11.13	+	2.61	8.90E-03	2.05
regulation of MAPK cascade (GO:0043408)	regulation of protein metabolic process (GO:0051246)	2475	30	12.03	+	2.49	1.37E-02	1.86
regulation of MAPK cascade (GO:0043408)	regulation of protein modification process (GO:0031399)	1631	23	7.93	+	2.9	2.66E-02	1.58
regulation of MAPK cascade (GO:0043408)	regulation of phosphorylation (GO:0042325)	1416	21	6.88	+	3.05	3.58E-02	1.45
regulation of MAPK cascade (GO:0043408)	regulation of MAPK cascade (GO:0043408)	670	14	3.26	+	4.3	4.23E-02	1.37
regulation of MAPK cascade (GO:0043408)	regulation of protein phosphorylation (GO:0001932)	1313	20	6.38	+	3.13	4.25E-02	1.37
regulation of multicellular organismal development (GO:2000026)	regulation of multicellular organismal development (GO:2000026)	1841	26	8.95	+	2.91	5.01E-03	2.30
regulation of multicellular organismal development (GO:2000026)	regulation of multicellular organismal process (GO:0051239)	2747	31	13.35	+	2.32	3.90E-02	1.41
response to retinoic acid (GO:0032526)	response to lipid (GO:0033993)	584	15	2.84	+	5.28	1.46E-03	2.84
response to retinoic acid (GO:0032526)	response to retinoic acid (GO:0032526)	77	6	0.37	+	16.03	2.08E-02	1.68
somatic stem cell population maintenance (GO:0035019)	stem cell population maintenance (GO:0019827)	135	8	0.66	+	12.19	3.23E-03	2.49
somatic stem cell population maintenance (GO:0035019)	maintenance of cell number (GO:0098727)	139	8	0.68	+	11.84	4.02E-03	2.40
somatic stem cell population maintenance (GO:0035019)	somatic stem cell population maintenance (GO:0035019)	48	5	0.23	+	21.43	3.70E-02	1.43
transcription, DNA-templated (GO:0006351)	organic cyclic compound metabolic process (GO:1901360)	3913	40	19.02	+	2.1	1.10E-02	1.96
transcription, DNA-templated (GO:0006351)	heterocycle metabolic process (GO:0046483)	3671	38	17.84	+	2.13	1.72E-02	1.76
transcription, DNA-templated (GO:0006351)	nucleobase-containing compound metabolic process (GO:0006139)	3544	37	17.22	+	2.15	2.05E-02	1.69
transcription, DNA-templated (GO:0006351)	cellular aromatic compound metabolic process (GO:0006725)	3725	38	18.1	+	2.1	2.46E-02	1.61
transcription, DNA-templated (GO:0006351)	macromolecule biosynthetic process (GO:0009059)	2712	31	13.18	+	2.35	2.99E-02	1.52
transcription, DNA-templated (GO:0006351)	transcription, DNA-templated (GO:0006351)	1903	25	9.25	+	2.7	3.19E-02	1.50
transcription, DNA-templated (GO:0006351)	nucleic acid-templated transcription (GO:0097659)	1904	25	9.25	+	2.7	3.22E-02	1.49
transcription, DNA-templated (GO:0006351)	RNA biosynthetic process (GO:0032774)	1907	25	9.27	+	2.7	3.31E-02	1.48
transcription, DNA-templated (GO:0006351)	organic cyclic compound biosynthetic process (GO:1901362)	2317	28	11.26	+	2.49	3.62E-02	1.44

transcription, DNA-templated (GO:0006351)	heterocycle biosynthetic process (GO:0018130)	2205	27	10.72	+	2.52	4.43E-02	1.35
transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	enzyme linked receptor protein signaling pathway (GO:0007167)	530	15	2.58	+	5.82	4.21E-04	3.38
transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	181	9	0.88	+	10.23	2.53E-03	2.60
	regulation of cell proliferation (GO:0042127)	1555	25	7.56	+	3.31	7.77E-04	3.11
	neurogenesis (GO:0022008)	1586	23	7.71	+	2.98	1.66E-02	1.78
	regulation of cell death (GO:0010941)	1600	23	7.78	+	2.96	1.93E-02	1.71
	negative regulation of gene expression (GO:0010629)	1488	22	7.23	+	3.04	2.13E-02	1.67

Table A.6 Gene Ontology terms enriched in the Class II (Silenced) ESC Super Enhancer associated gene set.

Gene Ontology (GO) was performed using the PANTHERDB over-representation test. P-values were adjusted for multiple hypothesis testing using the Bonferroni method.

Parent	GO biological process complete	Mus musculus - REFLIST (22221)	Class II (79)	Class II (expected)	Class II (over/under)	Class II (fold Enrichment)	Class II (P-value)	Class II (-log10(P-value))
chordate embryonic development (GO:0043009)	chordate embryonic development (GO:0043009)	698	14	2.48	+	5.64	1.40E-03	2.85
chordate embryonic development (GO:0043009)	embryo development ending in birth or egg hatching (GO:0009792)	707	14	2.51	+	5.57	1.64E-03	2.79
chordate embryonic development (GO:0043009)	embryo development (GO:0009790)	1056	16	3.75	+	4.26	7.05E-03	2.15
epithelial tube morphogenesis (GO:0060562)	epithelium development (GO:0060429)	987	19	3.51	+	5.41	1.15E-05	4.94
epithelial tube morphogenesis (GO:0060562)	multicellular organism development (GO:0007275)	4506	40	16.02	+	2.5	1.66E-05	4.78
epithelial tube morphogenesis (GO:0060562)	anatomical structure development (GO:0048856)	4795	41	17.05	+	2.41	2.77E-05	4.56
epithelial tube morphogenesis (GO:0060562)	developmental process (GO:0032502)	5122	42	18.21	+	2.31	5.53E-05	4.26
epithelial tube morphogenesis (GO:0060562)	tube development (GO:0035295)	627	14	2.23	+	6.28	3.80E-04	3.42
epithelial tube morphogenesis (GO:0060562)	anatomical structure morphogenesis (GO:0009653)	2031	24	7.22	+	3.32	7.31E-04	3.14
epithelial tube morphogenesis (GO:0060562)	morphogenesis of an epithelium (GO:0002009)	481	12	1.71	+	7.02	1.20E-03	2.92
epithelial tube morphogenesis (GO:0060562)	tissue morphogenesis (GO:0048729)	583	13	2.07	+	6.27	1.24E-03	2.91
epithelial tube morphogenesis (GO:0060562)	tissue development (GO:0009888)	1522	20	5.41	+	3.7	2.22E-03	2.65
epithelial tube morphogenesis (GO:0060562)	epithelial tube morphogenesis (GO:0060562)	347	10	1.23	+	8.11	3.95E-03	2.40
epithelial tube morphogenesis (GO:0060562)	tube morphogenesis (GO:0035239)	388	10	1.38	+	7.25	1.07E-02	1.97
epithelial tube morphogenesis (GO:0060562)	multicellular organismal process (GO:0032501)	7032	45	25	+	1.8	2.54E-02	1.60
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of transcription from RNA polymerase II promoter (GO:0006357)	1837	26	6.53	+	3.98	3.91E-06	5.41
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	cellular nitrogen compound biosynthetic process (GO:0044271)	2579	29	9.17	+	3.16	5.67E-05	4.25
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of macromolecule biosynthetic process (GO:0010556)	3580	33	12.73	+	2.59	4.51E-04	3.35
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of cellular biosynthetic process (GO:0031327)	1405	20	5	+	4	6.12E-04	3.21
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of gene expression (GO:0010468)	3837	34	13.64	+	2.49	6.71E-04	3.17
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of RNA metabolic process (GO:0051252)	3273	31	11.64	+	2.66	7.78E-04	3.11
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of transcription, DNA-templated (GO:0006355)	3094	30	11	+	2.73	8.32E-04	3.08
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of cellular macromolecule biosynthetic process (GO:2000112)	3478	32	12.36	+	2.59	8.50E-04	3.07
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of biosynthetic process (GO:0009890)	1435	20	5.1	+	3.92	8.62E-04	3.06
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of nucleic acid-templated transcription (GO:1903506)	3109	30	11.05	+	2.71	9.28E-04	3.03
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of RNA biosynthetic process (GO:2001141)	3115	30	11.07	+	2.71	9.68E-04	3.01
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of macromolecule biosynthetic process (GO:0010558)	1334	19	4.74	+	4.01	1.39E-03	2.86
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of gene expression (GO:0010629)	1488	20	5.29	+	3.78	1.55E-03	2.81
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of cellular biosynthetic process (GO:0031326)	3770	33	13.4	+	2.46	1.59E-03	2.80

negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of transcription, DNA-templated (GO:0045892)	1093	17	3.89	+	4.37	2.08E-03	2.68
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of biosynthetic process (GO:0009889)	3826	33	13.6	+	2.43	2.26E-03	2.65
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of nucleobase-containing compound metabolic process (GO:0019219)	3634	32	12.92	+	2.48	2.38E-03	2.62
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of nucleic acid-templated transcription (GO:1903507)	1109	17	3.94	+	4.31	2.55E-03	2.59
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of RNA biosynthetic process (GO:1902679)	1110	17	3.95	+	4.31	2.58E-03	2.59
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	1283	18	4.56	+	3.95	3.91E-03	2.41
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of macromolecule metabolic process (GO:0010605)	2276	24	8.09	+	2.97	5.96E-03	2.22
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of RNA metabolic process (GO:0051253)	1189	17	4.23	+	4.02	6.70E-03	2.17
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	796	14	2.83	+	4.95	6.71E-03	2.17
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of macromolecule metabolic process (GO:0060255)	5312	39	18.89	+	2.07	6.82E-03	2.17
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of nucleobase-containing compound metabolic process (GO:0045934)	1344	18	4.78	+	3.77	7.65E-03	2.12
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of cellular metabolic process (GO:0031324)	2316	24	8.23	+	2.91	8.16E-03	2.09
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of nitrogen compound metabolic process (GO:0051172)	2154	23	7.66	+	3	8.87E-03	2.05
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	negative regulation of metabolic process (GO:0009892)	2522	25	8.97	+	2.79	1.02E-02	1.99
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of nitrogen compound metabolic process (GO:0051171)	5169	38	18.38	+	2.07	1.04E-02	1.98
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	regulation of primary metabolic process (GO:0080090)	5294	38	18.82	+	2.02	1.94E-02	1.71
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	positive regulation of RNA metabolic process (GO:0051254)	1477	19	5.25	+	3.62	6.53E-03	2.19
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	positive regulation of nucleic acid-templated transcription (GO:1903508)	1404	18	4.99	+	3.61	1.43E-02	1.84
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	positive regulation of transcription, DNA-templated (GO:0045893)	1404	18	4.99	+	3.61	1.43E-02	1.84
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	positive regulation of RNA biosynthetic process (GO:1902680)	1405	18	5	+	3.6	1.44E-02	1.84
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	1117	16	3.97	+	4.03	1.46E-02	1.84
transcription, DNA-templated (GO:0006351)	nucleobase-containing compound biosynthetic process (GO:0034654)	2139	29	7.6	+	3.81	7.19E-07	6.14
transcription, DNA-templated (GO:0006351)	heterocycle biosynthetic process (GO:0018130)	2205	29	7.84	+	3.7	1.48E-06	5.83
transcription, DNA-templated (GO:0006351)	transcription, DNA-templated (GO:0006351)	1903	27	6.77	+	3.99	1.51E-06	5.82
transcription, DNA-templated (GO:0006351)	nucleic acid-templated transcription (GO:0097659)	1904	27	6.77	+	3.99	1.53E-06	5.82
transcription, DNA-templated (GO:0006351)	RNA biosynthetic process (GO:0032774)	1907	27	6.78	+	3.98	1.58E-06	5.80
transcription, DNA-templated (GO:0006351)	aromatic compound biosynthetic process (GO:0019438)	2213	29	7.87	+	3.69	1.62E-06	5.79
transcription, DNA-templated (GO:0006351)	organic cyclic compound biosynthetic process (GO:1901362)	2317	29	8.24	+	3.52	4.77E-06	5.32
transcription, DNA-templated (GO:0006351)	biosynthetic process (GO:0009058)	3545	34	12.6	+	2.7	8.95E-05	4.05
transcription, DNA-templated (GO:0006351)	cellular biosynthetic process (GO:0044249)	3395	33	12.07	+	2.73	1.21E-04	3.92
transcription, DNA-templated (GO:0006351)	organic substance biosynthetic process (GO:1901576)	3479	33	12.37	+	2.67	2.23E-04	3.65
transcription, DNA-templated (GO:0006351)	nucleobase-containing compound metabolic process (GO:0006139)	3544	33	12.6	+	2.62	3.52E-04	3.45
transcription, DNA-templated (GO:0006351)	RNA metabolic process (GO:0016070)	2662	28	9.46	+	2.96	4.92E-04	3.31
transcription, DNA-templated (GO:0006351)	cellular macromolecule biosynthetic process (GO:0034645)	2677	28	9.52	+	2.94	5.55E-04	3.26
transcription, DNA-templated (GO:0006351)	macromolecule biosynthetic process (GO:0009059)	2712	28	9.64	+	2.9	7.32E-04	3.14
transcription, DNA-templated (GO:0006351)	nucleic acid metabolic process (GO:0090304)	3085	30	10.97	+	2.74	7.79E-04	3.11
transcription, DNA-templated (GO:0006351)	heterocycle metabolic process (GO:0046483)	3671	33	13.05	+	2.53	8.34E-04	3.08
transcription, DNA-templated (GO:0006351)	cellular aromatic compound metabolic process (GO:0006725)	3725	33	13.24	+	2.49	1.19E-03	2.92
transcription, DNA-templated (GO:0006351)	organic cyclic compound metabolic process (GO:1901360)	3913	33	13.91	+	2.37	3.85E-03	2.41
transcription, DNA-templated (GO:0006351)	cellular metabolic process (GO:0044237)	7124	47	25.33	+	1.86	4.12E-03	2.39
transcription, DNA-templated (GO:0006351)	cellular nitrogen compound metabolic process (GO:0034641)	4092	33	14.55	+	2.27	1.10E-02	1.96

transcription, DNA-templated (GO:0006351)	gene expression (GO:0010467)	2965	27	10.54	+	2.56	1.69E-02	1.77
transcription, DNA-templated (GO:0006351)	organic substance metabolic process (GO:0071704)	7702	47	27.38	+	1.72	4.89E-02	1.31
	system development (GO:0048731)	3943	35	14.02	+	2.5	3.60E-04	3.44
	anatomical structure formation involved in morphogenesis (GO:0048646)	851	15	3.03	+	4.96	2.49E-03	2.60
	cellular developmental process (GO:0048869)	3496	29	12.43	+	2.33	3.97E-02	1.40

Table A.7 Mean-rank gene set test of Class I and II ESC Super Enhancer associated genes.

The pathways and corresponding genes were sourced from the CPDB integrative pathway database. The results are represented as ranks. The smaller the rank the more likely the pathway is enriched for differentially expressed genes between ESCs and EpiSCs. The Rank Sum column contains the sum of the ranks from both classes. The table displays only those pathways with rank ≤ 0.5 in at least one of the classes of Super Enhancer associated genes.

Pathway	Class I rank	Class II rank	Rank Sum
PluriNetWork	0.00	0.00	0.00
Hippo signaling pathway - Mus musculus (mouse)	0.06	0.06	0.12
Insulin Signaling	0.10	0.03	0.12
Assembly of collagen fibrils and other multimeric structures	0.10	0.09	0.19
Collagen formation	0.10	0.09	0.19
Integrin cell surface interactions	0.10	0.14	0.24
Transmembrane transport of small molecules	0.10	0.15	0.24
Metabolism of lipids and lipoproteins	0.03	0.23	0.26
Hemostasis	0.05	0.23	0.28
Focal adhesion - Mus musculus (mouse)	0.05	0.23	0.28
p53 signaling pathway - Mus musculus (mouse)	0.05	0.23	0.28
Cell cycle	0.05	0.23	0.28
Extracellular matrix organization	0.03	0.26	0.29
Leukocyte transendothelial migration - Mus musculus (mouse)	0.10	0.23	0.32
Type I hemidesmosome assembly	0.10	0.23	0.32
Cell junction organization	0.10	0.23	0.32
Cell-Cell communication	0.10	0.23	0.32
Platelet degranulation	0.10	0.23	0.32
Response to elevated platelet cytosolic Ca ²⁺	0.10	0.23	0.32
Platelet activation, signaling and aggregation	0.10	0.23	0.32
Fatty acid, triacylglycerol, and ketone body metabolism	0.18	0.23	0.41
Translation	0.23	0.23	0.46
Metabolism of proteins	0.26	0.23	0.49
Developmental Biology	0.02	1.00	1.02
Wnt signaling pathway - Mus musculus (mouse)	0.02	1.00	1.02
Interferon Signaling	0.03	1.00	1.03
Jak-STAT signaling pathway - Mus musculus (mouse)	0.03	1.00	1.03
Regulation of actin cytoskeleton - Mus musculus (mouse)	0.05	1.00	1.05
Translation Factors	0.05	1.00	1.05
Translation initiation complex formation	0.05	1.00	1.05
TCR signaling	0.06	1.00	1.06
Axon guidance	0.08	1.00	1.08
Cytokine Signaling in Immune system	0.08	1.00	1.08
Laminin interactions	1.00	0.09	1.09
Alzheimer,s disease - Mus musculus (mouse)	0.10	1.00	1.10
Maturity onset diabetes of the young - Mus musculus (mouse)	0.10	1.00	1.10
Glutathione metabolism	0.10	1.00	1.10
One carbon metabolism and related pathways	0.10	1.00	1.10
Glutathione and one carbon metabolism	0.10	1.00	1.10
Metabolism of water-soluble vitamins and cofactors	0.10	1.00	1.10
Metabolism of vitamins and cofactors	0.10	1.00	1.10
Vitamin C (ascorbate) metabolism	0.10	1.00	1.10
Defective TCN2 causes hereditary megaloblastic anemia	0.10	1.00	1.10
Defective AMN causes hereditary megaloblastic anemia 1	0.10	1.00	1.10
Defective CUBN causes hereditary megaloblastic anemia 1	0.10	1.00	1.10
Defective LMBRD1 causes methylmalonic aciduria and homocystinuria type cblF	0.10	1.00	1.10
Defective GIF causes intrinsic factor deficiency	0.10	1.00	1.10
Defective MMACHC causes methylmalonic aciduria and homocystinuria type cblC	0.10	1.00	1.10
Defective MMADHC causes methylmalonic aciduria and homocystinuria type cblD	0.10	1.00	1.10
Defective MTRR causes methylmalonic aciduria and homocystinuria type cblE	0.10	1.00	1.10
Defective MTR causes methylmalonic aciduria and homocystinuria type cblG	0.10	1.00	1.10
Defective MMAB causes methylmalonic aciduria type cblB	0.10	1.00	1.10
Defective MMAA causes methylmalonic aciduria type cblA	0.10	1.00	1.10
Defective MUT causes methylmalonic aciduria mut type	0.10	1.00	1.10
Defective CD320 causes methylmalonic aciduria	0.10	1.00	1.10

Defects in cobalamin (B12) metabolism	0.10	1.00	1.10
Defective HLCS causes multiple carboxylase deficiency	0.10	1.00	1.10
Defective BTD causes biotinidase deficiency	0.10	1.00	1.10
Defects in biotin (B7n) metabolism	0.10	1.00	1.10
Glucose transport	0.10	1.00	1.10
Defects in vitamin and cofactor metabolism	0.10	1.00	1.10
Facilitative Na ⁺ -independent glucose transporters	0.10	1.00	1.10
Hexose transport	0.10	1.00	1.10
TGF-beta signaling pathway - Mus musculus (mouse)	0.10	1.00	1.10
Pyrimidine salvage reactions	0.12	1.00	1.12
Pyrimidine metabolism	0.12	1.00	1.12
Metabolism of nucleotides	0.12	1.00	1.12
salvage pathways of pyrimidine ribonucleotides	0.12	1.00	1.12
IL-5 Signaling Pathway	0.12	1.00	1.12
Protein digestion and absorption - Mus musculus (mouse)	1.00	0.14	1.14
HIF-1 signaling pathway - Mus musculus (mouse)	0.18	1.00	1.18
Type II diabetes mellitus - Mus musculus (mouse)	0.18	1.00	1.18
Folate biosynthesis - Mus musculus (mouse)	0.18	1.00	1.18
Steroid hormone biosynthesis - Mus musculus (mouse)	0.18	1.00	1.18
Insulin secretion - Mus musculus (mouse)	0.18	1.00	1.18
Bile secretion - Mus musculus (mouse)	0.18	1.00	1.18
Non-alcoholic fatty liver disease (NAFLD) - Mus musculus (mouse)	0.18	1.00	1.18
Type II interferon signaling (IFNG)	0.18	1.00	1.18
Endochondral Ossification	0.18	1.00	1.18
Myometrial Relaxation and Contraction Pathways	0.18	1.00	1.18
Adipocytokine signaling pathway - Mus musculus (mouse)	0.18	1.00	1.18
Other glycan degradation - Mus musculus (mouse)	1.00	0.23	1.23
Amoebiasis - Mus musculus (mouse)	1.00	0.23	1.23
Hedgehog signaling pathway - Mus musculus (mouse)	1.00	0.23	1.23
SRF and miRs in Smooth Muscle Differentiation and Proliferation	1.00	0.23	1.23
3' -UTR-mediated translational regulation	0.23	1.00	1.23
Cell surface interactions at the vascular wall	0.24	1.00	1.24
Degradation of the extracellular matrix	1.00	0.26	1.26
Immune System	0.31	1.00	1.31
TNF signaling pathway - Mus musculus (mouse)	0.33	1.00	1.33
Apoptosis	0.33	1.00	1.33
Metabolism of carbohydrates	0.34	1.00	1.34
Cellular Senescence	0.34	1.00	1.34
Cellular responses to stress	0.34	1.00	1.34
Melanoma - Mus musculus (mouse)	0.42	1.00	1.42
Insulin signaling pathway - Mus musculus (mouse)	0.42	1.00	1.42
Signaling by ERBB2	0.42	1.00	1.42
Signaling by ERBB4	0.42	1.00	1.42
IGF1R signaling cascade	0.42	1.00	1.42
FRS2-mediated cascade	0.42	1.00	1.42
SHC-mediated cascade	0.42	1.00	1.42
Downstream signaling of activated FGFR	0.42	1.00	1.42
Signaling by FGFR mutants	0.42	1.00	1.42
IRS-related events triggered by IGF1R	0.42	1.00	1.42
Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)	0.42	1.00	1.42
IRS-mediated signalling	0.42	1.00	1.42
IRS-related events	0.42	1.00	1.42
Signaling by Leptin	0.42	1.00	1.42
Insulin receptor signalling cascade	0.42	1.00	1.42
Signaling by Insulin receptor	0.42	1.00	1.42
DAP12 signaling	0.42	1.00	1.42
DAP12 interactions	0.42	1.00	1.42
Signaling by Interleukins	0.42	1.00	1.42
NGF signalling via TRKA from the plasma membrane	0.42	1.00	1.42
Downstream signal transduction	0.42	1.00	1.42
Signaling by PDGF	0.42	1.00	1.42
Negative regulation of FGFR signaling	0.48	1.00	1.48
Fc epsilon receptor (FCER1) signaling	0.50	1.00	1.50
Innate Immune System	0.50	1.00	1.50
Downstream Signaling Events Of B Cell Receptor (BCR)	0.50	1.00	1.50
Signaling by the B Cell Receptor (BCR)	0.50	1.00	1.50
Signalling by NGF	0.50	1.00	1.50

Table A.8 The differential expression of shRNA targets in Nishyama et al.'s ESC shRNA screen.

The $\log_2FC$ column contains the $\log_2(\text{fold-change})$ in differential expression between the target shRNA treated and control ESCs. Colour key: red = positive $\log_2(FC)$ (i.e. increases in expression upon shRNA inhibition), yellow = $-1 > \log_2(FC) > 0$. shRNA targets indicating a positive $\log_2(FC)$ were excluded from analysis. The $-10 \times \log_{10}(\text{adj. } P\text{-value})$ column contains the associated P -value for differential expression. $P < 0.05$ equates to a $-10 \times \log_{10}(\text{adj. } P\text{-value}) > 13.31$. shRNA targets that were non-statistically significantly differentially expressed were also excluded from analysis.

Target	$\log_2FC$	$-10 \times \log_{10}(\text{adj. } P\text{-value})$
Dmrt2	0.52	13.31
Ncl	0.37	15.28
Batf3	0.24	55.80
Gabpa	0.19	36.87
4932441K18Rik	0.00	0.05
Foxd3	-0.08	3.57
Jarid2	-0.11	9.92
Eomes	-0.13	0.12
Sox1	-0.15	0.66
Fem1b	-0.15	22.63
Mkx1	-0.19	52.88
Foxp1	-0.35	42.78
Smad7	-0.37	48.09
Zscan10	-0.39	66.72
Etv1	-0.42	103.34
Sall1	-0.47	24.37
T	-0.48	17.25
Zfp57	-0.63	132.55
Dppa3	-0.82	162.83
Sox15	-0.85	372.68
Sfrs6	-0.86	294.45
Tcea3	-0.94	364.12
Pml	-0.99	383.00
Tcfap4	-0.99	288.42
Elf2	-1.07	66.68
Mtf2	-1.09	475.09
Fbxo15	-1.10	344.56
Aire	-1.13	203.33
Nanog	-1.13	356.56
Fgfbp1	-1.14	229.76
Inpp1	-1.14	207.65
Zfp42	-1.17	473.32
Zscan10	-1.20	328.72
Esrrb	-1.24	34.22
Smad4	-1.24	417.35
Grhl2	-1.26	202.26
Trim24	-1.27	393.49
Ash2l	-1.28	485.80
Trim28	-1.29	261.39
Zfp281	-1.34	671.87
Aff1	-1.45	611.52
Whsc2	-1.47	580.65
Foxg1	-1.48	408.96
Tbl1x	-1.58	777.83
Zic3	-1.59	485.91
Zfp292	-1.59	676.31
Prdm5	-1.60	741.54
Sall4	-1.61	128.12
Stat3	-1.67	627.96
Parp1	-1.76	722.41

<i>Rpap1</i>	-1.77	690.50
<i>Cdc5l</i>	-1.86	724.13
<i>Gli1</i>	-1.87	603.35
<i>Eed</i>	-1.88	628.59
<i>Dmrt1</i>	-1.89	493.13
<i>Phc1</i>	-1.90	815.08
<i>Btbd14b</i>	-1.90	91.79
<i>Ncoa3</i>	-1.92	590.52
<i>Gbx2</i>	-1.93	551.01
<i>Cdyl</i>	-1.93	712.58
<i>Sub1</i>	-1.94	557.10
<i>4932441K18Rik</i>	-1.97	403.00
<i>Tcfap2c</i>	-1.99	575.68
<i>Myc</i>	-1.99	452.34
<i>Nanog</i>	-2.03	592.76
<i>Klf5</i>	-2.05	617.06
<i>Id1</i>	-2.12	706.30
<i>Sox2</i>	-2.12	386.03
<i>Otx2</i>	-2.13	547.35
<i>Mycl1</i>	-2.17	527.33
<i>Tle4</i>	-2.18	716.73
<i>Hmga1</i>	-2.20	411.90
<i>Utf1</i>	-2.21	870.28
<i>Suz12</i>	-2.28	904.77
<i>Pdlim1</i>	-2.28	698.50
<i>Mybl2</i>	-2.29	127.37
<i>Tlk1</i>	-2.31	776.79
<i>Nr6a1</i>	-2.36	179.35
<i>Tcfcp2l1</i>	-2.36	727.74
<i>Tcl1</i>	-2.36	551.76
<i>Pou5f1-N (48hr)</i>	-2.38	472.07
<i>Nr5a2</i>	-2.39	435.37
<i>Ctbp2</i>	-2.49	971.56
<i>Mbd3</i>	-2.54	767.35
<i>Tbx3</i>	-2.55	692.66
<i>Jmjd1a</i>	-2.55	702.26
<i>Trp53</i>	-2.56	857.64
<i>Hells</i>	-2.57	959.59
<i>Cdyl2</i>	-2.60	919.42
<i>Gtf2i</i>	-2.61	816.19
<i>Dnmt3l</i>	-2.81	570.86
<i>Dnmt3b</i>	-2.82	659.32
<i>Dppa2</i>	-2.84	943.93
<i>Pou5f1-N (96hr)</i>	-2.85	555.61
<i>Lin28</i>	-2.88	752.20
<i>Mycn</i>	-3.00	747.01
<i>Ell2</i>	-3.08	911.56
<i>Klf2</i>	-3.14	751.94
<i>Klf4</i>	-3.19	776.46
<i>Rest</i>	-3.22	853.66
<i>Jmjd2c</i>	-3.24	112.52
<i>Sap30</i>	-3.31	877.64
<i>Nr0b1</i>	-3.38	701.59
<i>Pou5f1-N (72 hr)</i>	-3.41	632.25
<i>Nr0b1</i>	-3.57	730.02
<i>Ebf1</i>	-3.77	327.12
<i>Dppa5a</i>	-3.95	892.14
<i>Ctnnb1</i>	-3.95	638.31
<i>Pou5f1-B</i>	-4.82	791.66

1.2 Code

Code pertaining to the computational methods used in this thesis can be found at:

<https://github.com/emmabell42/PhD>

1.2.1 Sequencing data processing

1.2.1.1 Single-end sequencing reads

```
#####
##
## This script processes single-end sequencing reads. Briefly, it downloads SRA
## files from the NCBI, converts them to fastq files, aligns them to the mm9
## genome using Bowtie2, and finally merges, sorts, and subsets the resulting
## bam files. Subsetting the bam files dramatically reduces their processing
## time in R. The subsetting bam files were not used for peak-calling.
##
#####
#
#####
#
## Download SRA files. This step
## requires a text file listing the file
## paths on separate lines.
#
#####
#
#Create the text file 'toDownload.txt' listing the paths to SRA files.
#!/bin/bash
baseDir="ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/"
while read LINE;
do
    wget -r -np -k -nH -R robots.txt $baseDir$LINE
done < toDownload.txt
#####
#
## Convert SRA to fastq format with
## Fastq Dump
#
#####
mkdir ./fastq
find . -name *.sra | while read LINE;
do
/data/seqtools/sratoolkit.2.4.2-centos_linux64/bin/fastq-dump.2.4.2 $LINE "`basename "$LINE"
.sra`.fastq";
done
mv *.fastq ./fastq/
#####
#
## Assess read quality with FastQC
#
#####
mkdir ./fastqc
for file in ./fastq/*.fastq;
do
/data/seqtools/FastQC/fastqc -o ./fastqc "$file";
done;
#
## Create the fasta file 'overrepresented.fa' of the overrepresented sequences found by FastQC
#
nice R
library(XML)
library(GenomicRanges)
library(ShortRead)
overrepresented <- c()
setwd(paste0(getwd(), "/fastqc"))
htmlFiles <- list.files()[grep(".html", list.files())]
for(i in 1:length(htmlFiles)){
tables <- readHTMLTable(htmlFiles[i])
contents <- unlist(lapply(tables,function(x) "Possible Source" %in% colnames(x)))
lengthContents <- length(grep(TRUE,contents))
if(lengthContents>0){
tmp <- as.character(tables[[grep(TRUE,contents)]][,1])
overrepresented <- c(overrepresented,tmp)
}
}
```

```

}
overrepresented <- overrepresented[!duplicated(overrepresented)]
names(overrepresented) <- paste0("Overrepresented",1:length(overrepresented))
overrepresented <- DNASTringSet(overrepresented)
writeFasta(overrepresented,"overrepresented.fa")
q()
n
#####
#
## Trim fastq files with Trimmomatic.
## This step requires a text file
## listing the base names of the fastq
## files on separate lines.
#
#####
mv ./fastqc/overrepresented.fa ./
mkdir fastq.trimmed
for file in ./fastq/*.fastq;
do
java -jar /data/seqtools/Trimmomatic-0.33/trimmomatic-0.33.jar SE -threads 12 -phred33 $file
./fastq.trimmed/"`basename "$file"`.fastq`_trimmed.fastq"
ILLUMINACLIP:overrepresented.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:20;
done
#####
#
## Assess read quality after trimming
## with FastQC
#
#####
mkdir ./fastq.trimmed.fastqc
for file in ./fastq.trimmed/*.fastq;
do
/data/seqtools/FastQC/fastqc -o ./fastq.trimmed.fastqc "$file";
done;
for file in ./fastq.trimmed.fastqc/*.zip;
do
unzip $file;
done;
mv *_fastqc ./fastq.trimmed.fastqc
find . -name *summary.txt > ./summaries.txt
#
## Read in FastQC HTML files and report any errors
#
nice R
summaries <- read.table("summaries.txt",stringsAsFactors=F)
summaries <- apply(summaries,1,function(x) substr(x,2,59))
toName <- c()
toName <- apply(summaries,1,function(x) substr(x,2,12))
summaryTable <- c()
for(i in 1:length(summaries)){
toRead <- paste0(getwd(),summaries[i])
summary <- read.table(toRead,sep="\t")
summaryTable <- rbind(summaryTable,summary)
}
summaryTable <- summaryTable[which(!summaryTable[,1]=="PASS"),]
table(summaryTable[,2])
q()
n
#####
#
## Align reads with Bowtie2. This step
## requires a text file listing the base
## names of the fastq files on separate
## lines.
#
#####
export BOWTIE2_INDEXES=/data/seqtools/genomes/mm9/bowtie2/
mkdir sam
mkdir bam
find . -name "*_trimmed.fastq" | while read filename;
do
/data/seqtools/bowtie2-2.2.3/bowtie2 -x mm9 -U $filename > ./sam/"`basename "$filename"
.trimmed.fastq`.sam" -p 8;
samtools view -bS ./sam/"`basename "$filename"`.trimmed.fastq`.sam" > ./bam/"`basename
"$filename"`.trimmed.fastq`.bam";
done
#####

```

```
#
## Merge bam files with Samtools Merge.
## This step requires a comma delimited
## file where the first column is the
## merged bam file name and the
## subsequent columns at the bam files
## to merge.
#
#####
#
## Requires a tab-delimited table containing the name and bams to merge.
#
cd bam
while read LINE;
do
cols=$( echo $LINE | awk '{print NF}')
echo $cols
if [ $cols -eq 2 ]
then
repl=`echo $LINE | cut -f 2 -d " "`
name=`echo $LINE | cut -f 1 -d " "`
cp $repl ../merged.bam/$name.bam
else
bam=`echo $LINE | cut -f 2- -d " "`
name=`echo $LINE | cut -f 1 -d " "`
samtools merge ../merged.bam/$name".bam" $bam
fi
done < toMerge.txt
#####
#
## Sort and subset bam files to regions
## of interest with Samtools Sort and
## Bedtools Intersect
#
#####
mkdir ./subset.bam
for FILE in ../merged.bam/*.bam
do
samtools sort -o $FILE -O bam -T tmp $FILE -@12
bedtools intersect -abam $FILE -b /data/emmabell42/resources/ESC_enhancers_plus20.bed >
./subset.bam/"`basename "$FILE" .merged.bam`.subset.bam" -sorted;
done
```

1.2.1.2 Paired end sequencing reads

```
#####
##
## This script processes paired-end sequencing reads. Briefly, it downloads SRA
## files from the NCBI, converts them to fastq files, aligns them to the mm9
## genome using Bowtie2, and finally merges, sorts, and subsets the resulting
## bam files. Subsetting the bam files dramatically reduces their processing
## time in R. The subsetting bam files were not used for peak-calling.
##
#####
#
#####
#
## Download SRA files. This step
## requires a text file listing the file
## paths on separate lines.
#
#####
#
#Create the text file 'toDownload.txt' listing the paths to SRA files.
#!/bin/bash
baseDir="ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/"
while read LINE;
do
wget -r -np -k -nH -R robots.txt $baseDir$LINE
done < toDownload.txt
#####
#
## Convert SRA to fastq format with
## Fastq Dump
#
#####
```

```

mkdir ./fastq
find . -name *.sra | while read LINE;
do
/data/seqtools/sratoolkit.2.4.2-centos_linux64/bin/fastq-dump.2.4.2 $LINE ``basename "$LINE"
.sra`.fastq";
done
mv *.fastq ./fastq/
#####
#
## Assess read quality with FastQC
#
#####
mkdir ./fastqc
for file in ./fastq/*.fastq;
do
/data/seqtools/FastQC/fastqc -o ./fastqc "$file";
done;
#
## Create the fasta file 'overrepresented.fa' of the overrepresented sequences found by FastQC
#
nice R
library(XML)
library(GenomicRanges)
library(ShortRead)
overrepresented <- c()
setwd(paste0(getwd(), "/fastqc"))
htmlFiles <- list.files()[grep(".html", list.files())]
for(i in 1:length(htmlFiles)){
tables <- readHTMLTable(htmlFiles[i])
contents <- unlist(lapply(tables,function(x) "Possible Source" %in% colnames(x)))
lengthContents <- length(grep(TRUE,contents))
if(lengthContents>0){
tmp <- as.character(tables[[grep(TRUE,contents)]][,1])
overrepresented <- c(overrepresented,tmp)
}
}
overrepresented <- overrepresented[!duplicated(overrepresented)]
names(overrepresented) <- paste0("Overrepresented",1:length(overrepresented))
overrepresented <- DNAStringSet(overrepresented)
writeFasta(overrepresented,"overrepresented.fa")
q()
n
#####
#
## Trim fastq files with Trimmomatic.
## This step requires a text file
## listing the base names of the fastq
## files on separate lines.
#
#####
mkdir fastq.trimmedPE
while read LINE;
do
java -jar /data/seqtools/Trimmomatic-0.33/trimmomatic-0.33.jar PE -phred33 -baseout
/fastq.trimmedPE/$LINE $LINE"_R1.fastq" $LINE"_R2.fastq"
ILLUMINACLIP:overrepresented.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:20;
done < files
#####
#
## Align reads with Bowtie2. This step
## requires a text file listing the base
## names of the fastq files on separate
## lines.
#
#####
mkdir samPE
mkdir bamPE
export BOWTIE2_INDEXES=/data/seqtools/genomes/mm9/bowtie2/
while read LINE;
do
/data/seqtools/bowtie2-2.2.3/bowtie2 -x mm9 -1 ./fastq.trimmedPE/$LINE"_1P" -2
./fastq.trimmedPE/$LINE"_2P" -S ./samPE/$LINE".sam" -p 8
samtools view -bS ./samPE/$LINE".sam" > ./bamPE/$LINE".bam";

```

```
done < files
#####
#
## Merge bam files with Samtools Merge.
## This step requires a comma delimited
## file where the first column is the
## merged bam file name and the
## subsequent columns at the bam files
## to merge.
#
#####
mkdir merged.bam
cd bamPE
while read LINE;
do
cols=$( echo $LINE | awk '{print NF}')
echo $cols
if [ $cols -eq 2 ]
then
repl=`echo $LINE | cut -f 2 -d ","`
name=`echo $LINE | cut -f 1 -d ","`
cp $repl ../merged.bam/$name.bam
else
bam=`echo $LINE | cut -f 2- -d ","`
name=`echo $LINE | cut -f 1 -d ","`
samtools merge ../merged.bam/$name".bam" $bam
fi
done < toMerge.txt
cd ..
#####
#
## Sort and subset bam files to regions
## of interest with Samtools Sort and
## Bedtools Intersect
#
#####
mkdir ./subset.bam
for FILE in ../merged.bam/*.bam
do
samtools sort -o $FILE -O bam -T tmp $FILE -@12
bedtools intersect -abam $FILE -b /data/emmabell42/resources/ESC_enhancers_plus20.bed >
./subset.bam/"`basename "$FILE" ".merged.bam".subset.bam" -sorted;
done
```

1.2.1.3 Generating coverage plots from bam files

```
#####
##
## This script generates plots of the average sequencing read coverage of a bam
## file at a set of genomic regions. It first reads the bam files into R,
## removes PCR duplicates, corrects for sequencing depth and normalises to
## input. Dr Ed Curry wrote the plotCoverage function. It calculates the binned
## average sequencing read coverage of an IRanges coverage object across a set
## of genomic regions and outputs a plot. Input variables are capitalised.
##
#####
#
#####
#
## Load the required packages and create
## the plotCoverage function
#
#####
nice R
require(ShortRead)
library(GenomicRanges)
library(GenomicAlignments)
plotCoverage <- function(cov,regions,precision=1000,normalize.widths=TRUE,col="black",lty=1,lwd=1,plotfile=NULL){

  trackVal <- array(0,dim=c(length(regions),precision))
  cat(paste(nrow(trackVal),"sets of regions provided\n"))
  if(length(col)<nrow(trackVal)) col <- rep(col,nrow(trackVal))
  if(length(lty)<nrow(trackVal)) lty <- rep(lty,nrow(trackVal))
  if(length(lwd)<nrow(trackVal)) lwd <- rep(lwd,nrow(trackVal))
```

```

if(!normalize.widths) refMid <- ceiling(precision/2)
for(i in 1:length(regions)){
  theseSpaces <- intersect(as.character(unique(regions[[i]]@seqnames)),names(cov))
  cat(paste("region set",i,"has",length(theseSpaces),"spaces \n"))
  nrangesByChr <- rep(NA,length(theseSpaces))
  for(k in 1:length(theseSpaces)){
    cat(paste("\t calculating average coverage
for",sum(regions[[i]]@seqnames==theseSpaces[k]),"regions on space",theseSpaces[k]," \n"))
    chrCov <-
Views(cov[[theseSpaces[k]]],ranges(regions[[i]]@seqnames==theseSpaces[k]))
    nrangesByChr[k] <- length(chrCov)
    for(j in 1:length(chrCov)){
      thisCov <- chrCov[[j]]
      if(normalize.widths){
        if(length(thisCov)<precision) thisCov <-
Rle(rep(as.numeric(thisCov),each=floor(precision/length(thisCov))))
        indices <- round((0:precision)*(length(thisCov)/precision))
        starts <- indices[1:(length(indices)-1)]+1
        ends <- indices[2:length(indices)]
        newvals <- mean(Views(thisCov,start=starts,end=ends))
        newvals[is.na(newvals)] <- 0
        trackVal[i,] <- trackVal[i,] + newvals
      }
      if(!normalize.widths){
        thisMid <- ceiling(length(thisCov)/2)
        refStart <- max(c(0,refMid-thisMid))+1
        refEnd <- min(c(precision,refMid+length(thisCov)-thisMid))
        thisStart <- max(c(0,thisMid-refMid))+1
        thisEnd <- min(c(length(thisCov),thisMid+precision-refMid))
        cat(paste("filling ref from",refStart,"to",refEnd,"with Cov
from",thisStart,"to",thisEnd,"midpoints:",refMid,thisMid,"precision:",precision," \n"))
        newvals <- as.numeric(thisCov)[thisStart:thisEnd]
        newvals[is.na(newvals)] <- 0
        trackVal[i,refStart:refEnd] <- trackVal[i,refStart:refEnd] + newvals
      }
    }
  }
  trackVal[i,] <- trackVal[i,]/sum(nrangesByChr)
}

cat("making plot and returning output \n")
if(!is.null(plotfile)) png(plotfile)
cat(paste("range of values:",range(trackVal)," \n"))
plot(trackVal[1,],lty=lty[1],col=col[1],lwd=lwd[1],xlab="relative position",ylab="averaged
coverage",type="l",ylim=range(trackVal,na.rm=TRUE))
if(length(regions)>1){
  for(i in 2:length(regions)){
    points(trackVal[i,],lty=lty[i],col=col[i],lwd=lwd[i],type="l")
  }
}
if(!is.null(plotfile)) dev.off()
trackVal
}

#####
#
## Read bed files of regions of interest
## into R.
#
#####

bedFiles <- c("BED1","BED2","BED3",..., "BEDN")
bed.gr <- paste0(bedToName, ".gr")
for(i in 1:length(bedFiles)){
  bed <- read.table(bedFiles[i],head=F,sep="\t",stringsAsFactors=T)
  colnames(bed)[1] <- "chr"
  colnames(bed)[2] <- "start"
  colnames(bed)[3] <- "end"
  assign(bedToName[i],bed)
  gr <- as(bed,"GRanges")
  assign(bed.gr[i],gr)
}
#####
#
## Read bam files into R. Remove PCR
## duplicates and normalise for

```

```
## sequencing depth.
#
#####
bamFiles <- list.files(recursive=T)[grep(".bam",list.files(recursive=T))]
covObjects <- c()
for(i in 1:length(bamFiles)){
  last <- length(strsplit(bamFiles[i],"/")[[1]])
  covObjects[i] <- strsplit(bamFiles[i],"/")[[1]][last]
}
covObjects <- gsub(".bam",".cov",covObjects)
for(i in 1:length(bamFiles)){
  cat(i,"Reading and removing duplicates from",bamFiles[i],"at",as.character(Sys.time()),"\n",sep=" ")
  tmp <- readGAlignments(bamFiles[i], param=ScanBamParam(what="qname"))
  tmp <- tmp[!duplicated(tmp@start)]
  cat(i,"Coverting",bamFiles[i],"to GRanges and Coverage object at",as.character(Sys.time()),"\n",sep=" ")
  tmp.ranges <- as(tmp, "GRanges")
  tmp.cov <- coverage(tmp.ranges)
  cat(i,"Normalising",bamFiles[i],"to sequencing depth at",as.character(Sys.time()),"\n",sep=" ")
  nReads <- length(tmp.ranges)
  normFactor <- 10^6/nReads
  normCoverage <- tmp.cov*normFactor
  assign(covObjects[i],normCoverage)
}
#####
#
## Normalise each IP to its control ChIP
## experiment. This requires a tab-
## delimited table where the first
## column lists the IP and the second
## column lists the control.
#
#####
inputs <- read.table(file="/PATH/TO/INPUTS.txt",sep="\t",head=T,stringsAsFactors=F)
for(i in 1:nrow(inputs)){
  ip <- get(inputs[i,1])
  ip <- ip+1
  input <- get(inputs[i,2])
  input <- input+1
  ip.norm <- ip/input
  assign(inputs[i,1],ip.norm)
}
#####
#
## Plot average sequencing read
## coverage at regions of interest.
#
#####
for(i in 1:nrow(inputs)){
  toPlot <- get(inputs[i,1])
  png(paste0(covObjects[i],".png"))
  plotCoverage(cov=toPlot,regions=list(REGIONS.gr),normalize.widths=TRUE,precision=1000,col=c(COLOURS))
  dev.off()
}
```

1.2.2 Microarray processing

1.2.2.1 Affymetrix arrays

```
#####
##
## This script loads Affymetrix microarray data into R, log2 transforms,
## quantile normalises, performs linear regression, and calls differentially
## expressed genes between two conditions. Input variables are capitalised.
##
#####
#
#####
#
## Load the relevant packages into R
#
#####
#
library(affy)
library(limma)
```

```
#####
#
## Load the data into R for processing
## as an Affy batch object
#
#####
rawdata <- ReadAffy()
targets <- readTargets("TARGETS",row.names="filename")
exprs <- pm(rawdata)
#####
#
## Log2 transform and quantile
## normalise the probe intensity values
#
#####
normdata <- rma(rawdata)
normexprs <- exprs(normdata)
#####
#
## Perform clustering analysis to
## check the arrays cluster as expected
## and boxplot to confirm normalised
## distribution
#
#####
plot(hclust(dist(t(norm))))
boxplot(log(normexprs,base=2))
#####
#
## Fit the probe intensities to a linear
## model
#
#####
design <- array(0,dim=c(X,Y))
colnames(design) <- c("COLNAMES")
design[X,Y] <- 1
cont <- makeContrasts(CONTRAST="CONDITION1-CONDITION2",levels=design)
fit <- lmFit(norm,design)
fit <- contrasts.fit(fit,cont)
fit <- eBayes(fit)
#####
#
## Annotate the probes with the names of
## their target genes
#
#####
library(annotate)
library(ARRAY)
fit$genes$Symbol <- getSYMBOL(fit$genes$ID,"ARRAY")
#####
#
## Output a table statistically
## evaluating differential expression
## between CONDITION1 and CONDITION2
#
#####
hits <- topTable(fit,number=nrow(norm))
```

1.2.2.2 Agilent arrays

```
#####
##
## This script loads Affymetrix microarray data into R, log2 transforms, quantile
## normalises, performs linear regression, and calls differentially expressed
## genes between two conditions. Input variables are capitalised.
##
#####
#
#####
#
## Load the relevant packages into R
#
#####
```

```

library(limma)
#####
#
## Load the data into R for processing
## as a lumi batch object
#
#####
targets <- readTargets("TARGETS.txt")
rg <- read.maimages(targets,source="agilent")
#####
#
## Background correct, log2 transform,
## and quantile normalise the probe
## intensity values
#
#####
bgcorrected <- backgroundCorrect(rg,method="normexp")
ma <- normalizeWithinArrays(bgcorrected,method="loess")
ave <- avereps(ma, ID=ma$genes$ProbeName)
#####
plot(hclust(dist(t(norm))))
boxplot(log(normexprs,base=2))
#####
#
## Fit the probe intensities to a linear
## model
#
#####
design <- array(0,dim=c(X,Y))
colnames(design) <- c("COLNAMES")
design[X,Y] <- 1
cont <- makeContrasts(CONTRAST="CONDITION1-CONDITION2",levels=design)
fit <- lmFit(norm,design)
fit <- contrasts.fit(fit,cont)
fit <- eBayes(fit)
#####
#
## Annotate the probes with the names of
## their target genes
#
#####
library(annotate)
library(ARRAY)
fit$genes$Symbol <- getSYMBOL(fit$genes$ID,"ARRAY")
#####
#
## Output a table statistically
## evaluating differential expression
## between CONDITION1 and CONDITION2
#
#####
hits <- topTable(fit,number=nrow(norm))

```

1.2.2.3 Illumina arrays

```

#####
##
## This script loads Illumina microarray data into R, log2 transforms, quantile
## normalises, performs linear regression, and calls differentially expressed
## genes between two conditions. Input variables are capitalised.
##
#####
#
#####
#
## Load the relevant packages into R
#
#####
#
library(lumi)
library(lumiMouseIDMapping)
library(lumiMouseAll.db)
library(limma)

```

```
#####
#
## Load the data into R for processing
## as a lumi batch object
#
#####
data <- read.table(file="PATH/TO/FILE",head=T)
lumi <- lumiR(file="PATH/TO/FILE",lib.mapping="lumiMouseIDMapping")
lumi.exprs <- exprs(lumi)
lumi.exprs[which(lumi.exprs<1)] <- 1
#####
#
## Log2 transform and quantile
## normalise the probe intensity values
#
#####
log <- log2(lumi.exprs)
norm <- lumiN(log,method="quantile")
#####
#
## Perform clustering analysis to
## check the arrays cluster as expected
## and boxplot to confirm normalised
## distribution
#
#####
plot(hclust(dist(t(norm))))
boxplot(log(normexprs,base=2))
#####
#
## Fit the probe intensities to a linear
## model
#
#####
design <- array(0,dim=c(X,Y))
colnames(design) <- c("COLNAMES")
design[X,Y] <- 1
cont <- makeContrasts(CONTRAST="CONDITION1-CONDITION2",levels=design)
fit <- lmFit(norm,design)
fit <- contrasts.fit(fit,cont)
fit <- eBayes(fit)
#####
#
## Annotate the probes with the names of
## their target genes
#
#####
probe.list <- rownames(exprs(lumi))
if (require(lumiMouseAll.db) & require(annotate)) {
  gene.symbol <- getSYMBOL(probe.list, 'lumiMouseAll.db')
  fit$genes <- data.frame(ID=probe.list, SYMBOL=gene.symbol)
}
fit$genes$SYMBOL <- pData(featureData(data))$SYMBOL
#####
#
## Output a table statistically
## evaluating differential expression
## between CONDITION1 and CONDITION2
#
#####
hits <- topTable(fit,number=nrow(norm))
```

1.2.2.4 *Compiling a table of t-statistics for each annotated gene in the mouse genome*

```
#####
##
## This script creates a table of statistics pertaining to the differential
## expression of all mouse genes listed in the Mouse Genome Index. Here it
## gathers t-statistics, but this script was also used to create tables of
## log2(fold-changes) and multiple hypothesis testing corrected P-values. For
## each microarray or RNA-seq experiment, it requires a table with columns for
```

```
## MGI gene symbol and t-statistic (e.g. Limma's topTable output). Input
## variables are capitalised.
##
#####
#
#####
#
## Read in the results tables from the
## differential expression analysis of
## your microarray/RNA-seq experiments.
#
#####
results <- c("results1","results2","results3,...,resultsN")
for(i in 1:length(results)){
tmp <- read.table(file=results[i],head=T,sep="\t",comment.char="",quote="")
assign(results[i],tmp)
}
#####
#
## Create a list of the unique gene
## symbols included in the differential
## expression experiments that are also
## included in the Mouse Genome Index.
#
#####
genes <- as.list(rep(NA,length(results)))
for(i in 1:length(results)){
j <- get(as.character(results[i]))
genes[[i]] <- as.character(j$SYMBOL)
}
genes <- as.character(unlist(genes))
genes.unique <- setdiff(genes,c(NA,""))
mgi.symbols <-
read.table(file="PATH/TO/GENESYMBOLS.txt",head=T,sep="\t",comment.char="",quote="",stringsAsFa
ctors=F)
genes.unique <- intersect(genes.unique,mgi.symbols$SYMBOL)
genes.unique <- sort(genes.unique)
#####
#
## Create a table of t-statistics for
## each gene in genes.unique.
#
#####
t.table <- array(NA,dim=c(length(genes.unique),length(results)))
rownames(t.table) <- genes.unique
colnames(t.table) <- results
for(i in 1:length(results)){
j <- get(as.character(results[i]))
thisStudy <- data.frame(j$SYMBOL,(if("t" %in% colnames(j)) j$t else j$adj.P.Val))
colnames(thisStudy) <- c("SYMBOL","t")
mapped <- which(unique(thisStudy$SYMBOL) %in% genes.unique)
tofill <- which(genes.unique %in% thisStudy$SYMBOL)
new.statistics <- thisStudy$t[!duplicated(thisStudy$SYMBOL)][mapped]
t.table[tofill,i] <- new.statistics[order(unique(thisStudy$SYMBOL)[mapped])]
}
#####
#
## Quantile normalise the table of t-
## statistics.
#
#####
library(preprocessCore)
t.table.norm <- normalize.quantiles(t.table)
colnames(t.table.norm) <- colnames(t.table)
rownames(t.table.norm) <- rownames(t.table)
boxplot(t.table.norm)
```

1.2.2.5 Meta-analyses

1.2.2.5.1 Fisher's method

```
#####
##
## This script generates a table of Fisher's combined P-values.
```

```
##
#####
#
#####
#
## Create the fishersMethod function
#
#####
fishersMethod <- function(x) ifelse(sum(!is.na(x))>0,pchisq(-2 *
sum(log(x[!is.na(x)])),df=2*length(x[!is.na(x)]),lower=FALSE),NA)
#####
#
## Create a table to populate with
## the fishersMethod output and fill it.
#
#####
fishers.table <- array(NA,dim=c(nrow(p.table.norm),1))
for(i in 1:nrow(p.table.norm)){
fishers.table[i,] <- fishersMethod(p.table.norm[i,])
}
rownames(fishers.table) <- rownames(p.table.norm)
colnames(fishers.table) <- "Combined P-value"
#####
#
## FDR correct the Fisher's P-values and
## add these to the table.
#
#####
table.p.adj <- as.matrix(p.adjust(fishers.table,method="fdr"))
fishers.table <- cbind(fishers.table,table.p.adj)
colnames(fishers.table) <- c("Combined P-value","Adj P-value")
fishers.table.order <- fishers.table[order(fishers.table[,2]),]
```

1.2.2.5.2 Rank products

```
#####
##
## This script performs Rank Products on a table of differential expression
## log2(fold-changes).
##
#####
#
#####
#
## Load the RankProd library.
#
#####
library(RankProd)
#####
#
## Create an origin vector to identify
## different experiments, and a class
## vector to identify different data
## classes.
#
#####
origin <- colnames(logFC.table)
logFCs.cl <- rep(1,length(results))
#####
#
## Run RPadvance. This will generate two
## tables where for each gene:
## - Table1 gives the statistical
## significance of downregulation;
## - Table2 gives the statistical
## significance of upregulation.
#
#####
RP.adv.out <-
RPadvance(logFC.table,logFCs.cl,origin,num.perm=100,logged=TRUE,gene.names=rownames(logFC.table),rand=123)
RP.genes <-
topGene(RP.adv.out,num.gene=nrow(logFC.table),method="pfp",logged=TRUE,logbase=2,gene.names=rownames(logFC.table))
```

```
down.genes <- RP.genes$Table1[RP.genes$Table1$pfpr <= 0.05,]
up.genes <- RP.genes$Table2[RP.genes$Table2$pfpr <= 0.05,]
```

1.2.2.6 Gene set enrichment

```
#####
##
## This script performs Gene Set Enrichment Analysis of candidate genes in
## ConsensusPathDB pathways. It requires a list of genes with an associated
## statistic and the ConsensusPathDB list of biological pathways with their
## genes identified by gene symbol. Input variables are capitalised.
##
#####
#
#####
#
## Read in ConsensusPathDB pathways and
## split the column containing the genes
## into a list
#
#####
pathways <- read.table(file="PATH/TO/CPDB_pathways_genes.tab",sep="\t",header=TRUE)
pathway.genes <- list()
for(i in 1:nrow(pathways)){
  pathway.genes[[i]] <- strsplit(as.character(pathways$hgnc_gene_ids[i]),split=",")[[1]]
}
#####
#
## Create a results table, populate it
## with the geneSetTest output and
## correct the P-values for multiple
## hypothesis testing
#
#####
pathway.enrichments <- array(NA,dim=c(nrow(pathways),1))
rownames(pathway.enrichments) <- pathways$pathway
library(limma)
for(j in 1:nrow(pathways)){
  pathway.enrichments[j,1] <- geneSetTest(index=which(GENES[,3] %in%
pathway.genes[[j]]),statistics=STATISTICS,ranks.only=FALSE,type="t",alternative="up")
}
pathway.enrichments[,1] <- p.adjust(pathway.enrichments[,1],method="BH")
```