IMPERIAL COLLEGE LONDON

MRC Clinical Sciences Centre

INTERPLAY BETWEEN POLYCOMB REPRESSION AND RNA POLYMERASE II REGULATION GENOME-WIDE IN MURINE ES CELLS

INÊS DE SANTIAGO DOMINGOS DE JESUS

Thesis submitted for the degree of Doctor of Philosophy

September 2011

Inês de Santiago Domingos de Jesus: *Interplay between Polycomb repression and RNA polymerase II regulation genome-wide in murine ES cells* , © September 2011

SUPERVISORS:
Dr. Ana Pombo

GROUP:
Genome Function Group

DEPARTMENT:
MRC Clinical Sciences Centre

TIME FRAME:
September 2011

ABSTRACT

Gene regulation is a complex process that involves recruitment of transcription machinery, chromatin modification, and co-transcriptional RNA processing. RNA polymerase II (RNAPII) transcribes protein-coding genes. It contains a carboxy-terminal domain (CTD) consisting of multiple repeats. Specific patterns of CTD modifications promote the recruitment of appropriate factors at different stages of the transcription cycle, integrating transcription with chromatin modification and co-transcriptional RNA processing. In pluripotent ES cells, Polycomb repressive complexes (PRCs) silence important developmental regulator genes.

I have investigated RNAPII regulation at PRC-target genes genome-wide in murine ES cells by mining ChIP-Seq and mRNA-Seq datasets, using computational approaches. PRC targets exhibit different expression and RNAPII states: Silent genes lacking RNAPII, silent genes associated with RNAPII which is phosphorylated on the serine 5 residue of the CTD (S5P) but lacking serine 2 phosphorylated residues (S2P), and expressed genes with both S5P and S2P. At silent PRC-target genes, S5P and PRCs co-occupy the same promoters and the complexes extend into coding regions in the absence of S2P, suggesting disruption of co-transcriptional RNA processing. Sequential ChIP demonstrates simultaneous association of PRC and RNAPII to the same chromatin. Most expressed PRC-target genes display CTD marks of activity, including presence of S2P throughout the gene body; however S2P and PRCs do not co-immunoprecipitate, suggesting that active and PRC-repression are two separate chromatin states, that occur in different cells of the ES population, or within different alleles in the same cell. Finally, I have developed a software, SeqGI, for the analysis and interpretation of sequencing data. SeqGI provides a GUI framework for the simultaneous visualisation and statistical evaluation of read profiles over genomic features.

This work provides valuable insights on the biology of ES cells, and elucidates combinations of different factors across the genome shedding light on different mechanisms of gene activation, poising and silencing.

DECLARATION

I declare that the thesis hereby submitted for the degree of Doctor of Philosophy at Imperial College London, presents my own research work and effort. Collaborative research and discussions are properly acknowledged in the text.

London, September 2011

Inês de Santiago Domingos
de Jesus

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to my family and friends.

"If your point of view is correct, each organism will be characteristized by a long number written in quadrucal (?) system with figures 1, 2, 3, 4 standing for different bases ... This would open a very exciting possibility of theoretical research based on combinatorix [sic] and the theory of numbers! ... I have a feeling this can be done. What do you think?"

— George Gamow, in a letter to Watson and Crick (1953)

ACKNOWLEDGMENTS

I would like to thank my supervisor, and friend, Ana Pombo, whose expertise, guidance, and understanding, added considerably to this four year journey. I am thankful for the confidence she placed in me from the beginning, and for her support and encouragement throughout the writing of this thesis.

I would like to thank Emily for her dedication to this project, for her persistence and for adding so much pleasure to my work during the time we were together.

Tom Carroll for all his support, both as friend and a colleague, and for immediately believing in the SeqGI project and having such an enthusiastic approach to it.

I would like to thank Sarah Teichmann and Daniel Hebenstreit who shared their ideas, and collective wisdom in meetings and discussions. I would also like to thank Madan Babu and Sarath Janga for their advice and discussions during the first year of my PhD. I am also thankful to Marie-France Sagot, for her invaluable assistance in the beginning of my PhD.

I would like to express my appreciation to Instituto Gulbenkian de Ciência and the organizers of PhD Program in Computational Biology (PDBC) and all my twelve colleagues, and to my sponsoring agency Fundação para a Ciência e Tecnologia. In particular, I am thankful to Jorge Carneiro, director of the PDBC, for his genuine interest in my research, and for the opportunity to embrace in the field of computational biology.

I thank all Genome Function group members with whom I so much enjoy working. I have had so many interesting discussions with all of you, Sheila, Kelly, Liron, Inês, Mita, Carmelo, Claudia, Kedar and João. Special thanks to Ines Castro, my friend, for her patient support, and for bringing a smile to work every day. Also special thanks

to my friend Liron who has been through this journey with me from the start, and with whom it has been a pleasure to work with. Also thanks to Carmelo Ferrai for his critical thinking and friendship.

To all contributors to this project, in particular those who work I present in this thesis, I offer thanks for advice, contribution and many stimulating discussions.

Last but not the least, to my family, Mum and Sister, for their endless support and encouragement, and for believing in me each step of the way.

CONTENTS

I INTRODUCTION 23

1 LITERATURE REVIEW AND THESIS OVERVIEW 25

1.1 Gene transcription and regulation . . . 26
1.1.1 Transcription initiation . . . 26
1.1.2 Transcription Elongation . . . 28
1.1.3 Transcription termination . . . 28
1.1.4 Core promoter elements and gene regulation . . . 29
1.1.5 Functions of gene clustering . . . 30
1.1.6 Epigenetic gene regulation . . . 31
1.1.7 Gene regulation in ES cells . . . 33
1.2 RNA Polymerase II and the C-terminal domain . . . 34
1.3 Integration of the CTD code with regulatory processes . . . 36
1.3.1 CTD serine 5 phosphorylation and transcription initiation . . . 36
1.3.2 CTD serine 2 phosphorylation and transcription elongation . . . 39
1.3.3 CTD serine 7 phosphorylation . . . 40
1.3.4 Other CTD modifications . . . 41
1.4 Transcriptional Priming and RNAPII pausing . . . 42
1.4.1 RNAPII pausing . . . 42
1.4.2 Control of RNAPII pausing . . . 43
1.4.3 RNAPII at silent Polycomb repressed genes . . . 44
1.5 The Polycomb repressive complexes . . . 44
1.5.1 Structure and function of Polycomb complexes . . . 44
1.5.2 Polycomb-mediated gene silencing . . . 46
1.5.3 Polycomb recruitment . . . 47
1.5.4 Chromatin bivalency . . . 48
1.6 Research aims and objectives . . . 50
1.7 Thesis outline . . . 50

II METHODS 53

2 METHODS 55

2.1 Experimental methods . . . 55
2.1.1 Cell culture . . . 55
2.1.2 Chromatin immunoprecipitation: single and sequential ChIP . . . 55
2.1.3 ChIP Cross-Linking Vs Native ChIP . . . 56
2.1.4 ChIP validation by qPCR . . . 56
2.1.5 RNA purification and gene expression analysis by RT-PCR . . . 57
2.1.6 Illumina library preparation and sequencing: ChIP-Seq and mRNA-Seq . . . 57
2.1.7 Antibody validation by ELISA . . . 57
2.1.8 Immuno-cryoFISH . . . 58
2.2 High-throughput sequencing and alignment . . . 58

2.2.1 ChIP-seq datasets . . . 58
2.2.2 mRNA-Seq datasets . . . 58
2.2.3 Alignment of sequenced reads . . . 59
2.3 Processing of genome-wide sequencing data . . . 61
2.3.1 Generation of bedGraph Files . . . 61
2.3.2 Computation of sequencing *depth of coverage* at genomic intervals of interest . . . 61
2.3.3 Data visualisation in the log scale . . . 61
2.3.4 Visualisation of the distribution pattern of ChIP-Seq reads in gene 5′ and 3′ regions . . . 62
2.3.5 FPKM calculation . . . 62
2.4 Analysis of Published annotation datasets . . . 62
2.4.1 Microarray datasets and gene expression analyses . . . 62
2.4.2 Other genome-wide annotation datasets . . . 63
2.5 Choice of a non-redundant genome . . . 65
2.6 Gene ontology analysis . . . 66
2.7 KEGG pathway analysis . . . 67
2.8 Classification scheme . . . 67
2.9 Hierarchical clustering . . . 68
2.10 Principal Component Analysis . . . 68
2.11 CpG and GC content analysis . . . 68
2.11.1 Calculation of percentage CpG and GC . . . 68
2.11.2 CpG O/E calculation and distribution pattern . . . 69
2.12 GSEA analysis . . . 69
2.13 SeqGI source code implementation . . . 69
2.13.1 Computation of overlapping genomic intervals via OGRe.py . 70
2.13.2 SeqGI graphical user interface . . . 70
2.13.3 Statistical analyses within SeqGI . . . 70

III RESULTS 71

3 GENOME-WIDE MAPPING OF RNAPII AND TRANSCRIPTOME ANALYSIS IN MURINE ES CELLS 73
3.1 Research Motivation . . . 73
3.2 State of the Art . . . 74
3.2.1 RNAPII CTD “code” . . . 74
3.2.2 RNAPII at bivalent genes in ES cells . . . 74
3.3 Results . . . 76
3.3.1 Generation of genome wide transcriptome data . . . 76
3.3.2 Generation of genome-wide profiles of RNAPII phosphorylation 78
3.3.3 Validation of the ChIP-Seq data by quantitative real-time PCR 81
3.3.4 RNAPII and transcriptome coverage of the mouse genome in ES cells . . . 84
3.3.5 RNAPII profiles at highly and minimally expressed genes . . 87
3.3.6 RNAPII is associated with both K4 monovalent promoters and K4+K27 bivalent promoters in ES cells . . . 90

3.3.7 RNAPII high-resolution Chip-Seq profiles reveal previously unknown patterns . . . 91
3.3.8 Comparison of RNAPII ChIP-Seq datasets with published datasets 95
3.3.9 Events associated with transcriptional elongation occur at a proportion of bivalent genes . . . 97
3.4 Discussion . . . 104
3.4.1 RNAPII configuration at active and inactive genes . . . 104
3.4.2 PRC-repressed RNAPII has an unusual combination of post-translational modifications in ES cells . . . 106
3.4.3 Transcription of bivalent genes . . . 107
4 POLYCOMB REPRESSION IN THE PRESENCE OF RNAPII 111
4.1 Research Motivation . . . 111
4.2 Results . . . 112
4.2.1 Genome-wide mapping of Polycomb and histone modification data in ES cells . . . 112
4.2.2 Chromatin and RNAPII patterns at Active, Bivalent and Inactive genes . . . 116
4.2.3 Classification RNAPII or PRCs ChIP-Seq signal . . . 119
4.2.4 Robustness of the classification . . . 128
4.2.5 Validation of the classification by comparing with previous classifications . . . 131
4.2.6 Characterization of PRC+ genes . . . 135
4.3 Discussion . . . 136
4.3.1 Distinguishing marked and bound genes from background . . . 137
4.3.2 Improvements on the classification of bivalent and PRC1-positive genes . . . 138
4.3.3 RNAPII at some PRC-repressed genes is transcriptionally active 139
5 GENE ACTIVITY IN THE PRESENCE OF POLYCOMB 141
5.1 Research Motivation . . . 141
5.2 Two alternative models for Polycomb regulation at "active" PRC-target genes . . . 141
5.3 Results . . . 143
5.3.1 Identification of three epigenetic states in mESCs . . . 143
5.3.2 Principal component analysis identifies clear separation between Active, PRC-repressed and Inactive classes . . . 146
5.3.3 Comparing PCA with the results from Hierarchical clustering 147
5.3.4 Subclassification of gene classes within the Active, PRC-repressed and Inactive branches of the HC tree . . . 149
5.3.5 Expression levels increase progressively from Inactive to Active groups . . . 150
5.3.6 Characteristic chromatin signatures: From Inactive to Active groups . . . 152
5.3.7 PRC-association is compatible with protein expression at PRCa genes . . . 154
5.3.8 Gene classes are distinguishable by promoter CpG content and DNA methylation . . . 155

5.3.9 PRCr gene promoters associated with broad S5P distribution and high CpG content . . . 157
5.3.10 Similar RNAPII and PRC enrichment patterns at PRCr genes . 159
5.3.11 PRC1, PRC2 and RNAPII physically co-associate at PRCr genes 161
5.3.12 Ring1B depletion affects the expression levels of PRCr, PRCi and PRCa genes . . . 163
5.3.13 PRCs functionally repress active developmental and metabolic genes in ES cells. . . . 166
5.3.14 PRCs physically co-associate with RNAPII-S5P but not S2P at active PRC-targets . . . 169
5.3.15 Pluripotent transcription factors are enriched at PRCa genes . 174
5.3.16 Single cell analysis of PRCa gene, Lefty2, by immuno-cryoFISH 178
5.4 Discussion . . . 179
5.4.1 The three major epigenetic states in ES cells: Active, PRC-repressed and Inactive . . . 180
5.4.2 CpG promoter structure and the regulation of gene expression 183
5.4.3 The merging of Polycomb- and DNA methylation-mediated gene silencing at PRCo genes . . . 184
5.4.4 Promoter-proximal pausing versus "poised" RNAPII at PRC-repressed genes . . . 185
5.4.5 Synergy between PRCs and RNAPII at PRCr genes . . . 186
5.4.6 Active PRC-repressed genes (PRCa, PRCi) as a read-out of heterogeneous ES cell populations . . . 187
5.4.7 Interactions between PRCs and the transcriptional network . . 189
6 SEQGI - A SOFTWARE TO ANALYSE CHIP-SEQ DATA 191
6.1 Research motivation . . . 191
6.2 State of the art . . . 191
6.2.1 Quantifying ChIP-Seq signal ("counts") *versus* identifying peaks 191
6.2.2 Current available packages to analyse and visualise ChIP-Seq count data . . . 192
6.3 SeqGI overview and implementation . . . 195
6.3.1 SeqGI workflow . . . 195
6.3.2 Layout of the GUI . . . 196
6.3.3 Identification of overlaps . . . 197
6.3.4 Types of Input files . . . 197
6.3.5 Flexibility of input file structure . . . 199
6.3.6 File sorting . . . 199
6.3.7 Intersection function . . . 200
6.4 SeqGI output and results . . . 200
6.4.1 Overlap scores are given in two types of files "Overlap" or "Sliding" . . . 200
6.4.2 Graphical visualisation . . . 202
6.4.3 Statistical testing . . . 203
6.4.4 Implementation . . . 204
6.5 Computing Overlaps between genomic intervals - OGRe script . . . 204

6.5.1 Algorithms to compute the overlap between two genomic feature sets . . . 204
6.5.2 Fjoin "sort merge-pass" algorithm . . . 205
6.5.3 OGRe improvements on Fjoin . . . 205
6.6 Sample workflow of SeqGI . . . 207
6.6.1 Installation and use . . . 207
6.6.2 Adding new files to a project . . . 207
6.6.3 Graphical representation . . . 208
6.7 Conclusion . . . 209

IV DISCUSSION 211

7 DISCUSSION 213
7.1 Thesis Overview . . . 213
7.2 From research objectives to research findings . . . 215
7.2.1 Bioinformatic analysis and software development . . . 215
7.2.2 A review of the CTD central dogma . . . 217
7.2.3 Interplay between RNAPII and Polycomb at PRC-repressed genes 218
7.2.4 Fluctuations between the canonical active and PRC-repressed states . . . 220
7.3 Future research directions . . . 221

V APPENDIX 257

BIBLIOGRAPHY 259

LIST OF FIGURES

Figure 1.1 Schematic representation of the transcriptional cycle in yeast. . 38
Figure 1.2 The current model of CTD phosphorylation patterns. 39
Figure 1.3 Polycomb repressive complexes. 46
Figure 1.4 Schematic representation of chromatin at bivalent genes in ES cells and upon differentiation. 49
Figure 2.1 Scheme of the methodology used to build a candidate gene set for downstream analyses. 66
Figure 3.1 Comparison of read counts between independent sequencing lanes of the mRNA library. 77
Figure 3.2 Inspection of mRNA-Seq profiles and levels across single genes. 78
Figure 3.3 Reproducibility of sequencing replicates of the same ChIP-Seq library. 81
Figure 3.4 ChIP-Seq and mRNA-Seq profiles across single genes. 83
Figure 3.5 Quality control qPCR for RNAPII and H3K36me3 ChIP samples. 84
Figure 3.6 Mapped RNAPII and mRNA-Seq reads are strongly biased towards annotated genic regions. 86
Figure 3.7 Genome-wide ChIP profiles at active and inactive genes. 89
Figure 3.8 RNAPII occupancy predicts mRNA expression. 90
Figure 3.9 Genome-wide RNAPII profiles at bivalent genes. 91
Figure 3.10 Occupancy of RNAPII, H3K36me3 and H3K4me3 at single genes. 94
Figure 3.11 Comparison of newly generated and published ChIP-Seq datasets for S5P and S2P forms of the RNAPII. 97
Figure 3.12 Comparison of expression levels between independent mRNA studies. 100
Figure 3.13 Expression of bivalent genes is detected in gene expression studies. 101
Figure 3.14 Expression of bivalent genes is positively associated with presence of 8WG16, S7P, S2P and H3K36me3. 103
Figure 4.1 Quality control qPCR for H2Aub1 and Ring1B ChIP samples. . 114
Figure 4.2 ChIP-Seq and mRNA-Seq profiles across single genes. 115
Figure 4.3 Genome-wide profiles of PRC related marks at most and least expressed and bivalent genes. 117
Figure 4.4 Pairwise correlation between levels of PRC markers at promoters. 118
Figure 4.5 Overview of the classification strategy. 120
Figure 4.6 Genome-wide average read profiles at RefSeq genes. 122
Figure 4.7 Curve fitting of RNAPII (S5P, 8WG16, S7P, S2P), H3K4me3, H3K27me3 and H2Aub1 data. 124
Figure 4.8 Curve fitting of the mock IP control distribution. 125
Figure 4.9 Example fitting of 3 models to H3K4me3 and S5P data sets. . . 126
Figure 4.10 Example of genes differentially classified between three approaches using three different window sizes (2kb, 1kb and 500bp). 130

Figure 4.11 Bivalent gene classifications in the current and previous studies. 132
Figure 4.12 Comparison of H3K4me3, H3K27me3 and H2Aub1 sequencing depth at the promoters of genes that differ between the current and previous classifications. 134
Figure 4.13 mRNA levels at PRC$^+$ genes encompass the spectrum of expression levels. 136
Figure 4.14 PRC-targets associate with different RNAPII modifications and expression levels. 136
Figure 5.1 Two alternative models for PRC-regulation at PRCa genes. . . 142
Figure 5.2 Hierarchical clustering of genes according to their combination of RNAPII and histone modifications. 144
Figure 5.3 PCA identifies three major groups within the data. 146
Figure 5.4 Comparison of Hierarchical clustering of genes with PCA results. 149
Figure 5.5 Levels of mRNA across Active, Inactive and PRC-cohorts. . . . 151
Figure 5.6 RT-PCR analyses confirm the expression status of genes from distinct PRC gene clusters. 152
Figure 5.7 Levels of mRNA, RNAPII, and PRC subunits across Active, Inactive and PRC-cohorts. 153
Figure 5.8 Gene expression at the protein level across Active, Inactive and PRC-cohorts. 154
Figure 5.9 Percentage of CpG and DNA methylation at Active, Inactive and PRC cohorts. 155
Figure 5.10 Genome-wide core H3 profiles at Active, Inactive and PRC cohorts. 157
Figure 5.11 Correlation between broad CpG content and S5P profiles at Active, Inactive and PRC cohorts. 159
Figure 5.12 S5P and PRCs show similar genome-wide ChIP profiles at PRCr genes. 160
Figure 5.13 Similar domains of enrichment of S5P, H3K27me3 and H2Aub1 at PRCr genes. 161
Figure 5.14 PRC1, PRC2 and RNAPII physically co-associate at PRCr genes. 163
Figure 5.15 PRCr, PRCi and PRCa genes are predominantly up-regulated after Ring1B-knockout. 165
Figure 5.16 Genes classified as positive for S5P at the TTS (S5PEnd+) are more likely to be de-repressed after Ring1B removal. 166
Figure 5.17 KEGG pathways significantly enriched in active PRC-target genes (PRCa cluster) 168
Figure 5.18 Changes in expression during differentiation for PRCa genes. . 169
Figure 5.19 Distinct combinations of RNAPII at each PRC class. 170
Figure 5.20 Simultaneous detection of repressive PRCs and activating RNAPII modifications at PRCi and PRCa genes. 171
Figure 5.21 PRCs physically co-associate with RNAPII-S5P but not S2P at active PRC-targets. 172
Figure 5.22 Strong positive correlation between S5P and H2Aub1 within PRCr and between S5P and S2P at Active and PRCa genes. . . 173

Figure 5.23 GSEA plot showing enrichment for Nanog targets in PRCa and Active groups. 177
Figure 5.24 Distribution of multiple/single TFs target genes among Active, PRC cohorts and Inactive groups of genes. 178
Figure 5.25 PRCs and elongating RNAPII-S2P are mutually exclusive at the *Lefty2* gene. 179
Figure 5.26 Model of gene regulatory states in ES cells. 181
Figure 6.1 A collection of SeqGI screen shots 195
Figure 6.2 The workflow of SeqGI. 196
Figure 6.3 Layout of SeqGI GUI. 197
Figure 6.4 Example of intersection between two genomic intervals. 200
Figure 6.5 Sample workflow of SeqGI - Adding new files to a project . . . 208
Figure 6.6 Sample workflow of SeqGI - Example plotting a Box Plot . . . 209
Figure 7.1 Schematic representation of average RNAPII and histone modification profiles at “Active”, “PRC-repressed” and “Inactive” states. 213
Figure 7.2 Model of the switch between PRC-repressed and the Active states in ES cells. 221

LIST OF TABLES

Table 1.1 CTD modifications and respective modifying enzymes and their effect on transcription. 42
Table 2.1 List of ChIP-Seq and mRNA-Seq datasets used in this thesis. . 60
Table 2.3 Sources of genomic annotations downloaded from selected published papers 64
Table 3.1 Number of mapped reads from mRNA-Seq dataset. 76
Table 3.2 List of ChIP-Seq datasets used in this study. 80
Table 3.3 Summary of read distribution for ChIP-Seq and mRNA-Seq datasets. 87
Table 3.4 List of published RNAPII ChIP-Seq datasets used in this study. 95
Table 3.5 List of mRNA datasets used. 98
Table 4.1 List of ChIP-Seq datasets used in this study. 113
Table 4.2 Parameters of the fitted distributions for different window sizes. 127
Table 4.3 Agreement between classifications using different window sizes. 129
Table 4.4 Number of overlapped windows excluded from each data set. 131
Table 5.1 Variance within the data explained by the principal components. 147
Table 5.2 Loadings of each principal component. 148
Table 5.3 Gene set enrichment analysis (GSEA) of TF targets in six clusters of genes. 176
Table 6.1 Comparison between SeqGI with available packages to visualise and analyse sequence count data. 194
Table 7.1 Summary of the three major epigenetic states in ES cells: “Active”, “PRC-repressed” and “Inactive”. 215

ACRONYMS

bioChIP biotinylation mediated ChIP

BIC Bayesian information criterion

BREu upstream TFIIB Recognition Element

ChIP-Seq Chromatin-immunoprecipitation followed by sequencing

CTD carboxy-terminal domain

CGI promoters CpG island promoters

DSIF DRB sensitivity-inducing factor

DPE Downstream Promoter Element

DCE Downstream Core Element

dKO double-knockout

ES embryonic stem

Ezh2 enhancer of zeste homolog 2

ELISA Enzyme Linked Immunosorbent Assay

GSEA Gene-set enrichment analysis

GO Gene Ontology

GUI graphical user interface

GTFs general transcription factors

Inr Initiator element

LIF recombinant leukemia inhibitory factor

lncRNAs long non-coding RNAs

miRNAs microRNAs

MTE Motif Ten Element

NELF negative elongation factor

nt nucleotides

NPCs neuronal progenitor cells

NGS Next-Generation Sequencing

mock IP DNA immunoprecipitated in the absence of antibody

ncRNAs non-coding RNAs

PCA principal component analysis

PIC preinitiation complex

P-TEFb positive transcription elongation factor b

RNAPII RNA polymerase II

re-ChIP Sequential ChIP

Suz12 suppressor of zeste 12

snRNAs small nuclear RNAs

TFIIH Transcription factor II H

TSS transcription start site

UCSC University of California Santa Cruz

NURF nucleosome remodeling factor

DSIF 5,6-dichloro-1-b-D-ribofuranoxylbenzimidazole sensivitiy-inducing factor

NELF negative elongation factor

PcG Polycomb group of proteins

Eed embryonic ectoderm development

Suz12 suppressor of zeste 12

Ezh2 enhancer of zeste homologue 2

PREs polycomb group responsive elements

Part I

INTRODUCTION

1

LITERATURE REVIEW AND THESIS OVERVIEW

Transcriptional regulation is a major mechanism to determine specific sophisticated responses to environmental conditions of a cell. The regulation of biological processes such as cell differentiation requires precise control of gene expression patterns. To understand the molecular mechanisms of gene regulation, comprehensive characterisation of functional elements associated to gene control is needed. Completion of the human genome sequence (Venter et al., 2001) together with the advent of Next-Generation Sequencing (NGS) technologies now provide the information and the possibility to study gene expression regulation on a global scale.

Here I review the mechanisms of gene regulation, particularly focusing in embryonic stem (ES) cells. I begin with an overview of the different levels of eukaryotic gene regulation and the complex molecular machinery that operates it. I then focus the regulatory processes important for ES cell genome regulation, including Polycomb-mediated gene repression.

1.1 GENE TRANSCRIPTION AND REGULATION

RNA polymerase II (RNAPII) transcribes all protein-coding genes and a variety of non-coding RNAs (ncRNAs) including small nuclear RNAs (snRNAs), microRNAs (miRNAs), and long non-coding RNAs (lncRNAs). In eukaryotes, regulation of gene expression is essential to drive the process of cellular specialization, leading to the creation of ~200 different cell types, each differentiated to serve a specialized function for which a specific gene expression profile is required, though they all carry the same genetic code. Gene regulation can occur at several steps, including transcription initiation, transcription elongation, mRNA processing, transport and translation. Gene transcription has been extensively studied in terms of its mechanism and regulation and it is the focus of this introduction.

Gene transcription is stimulated and regulated by *cis*-regulatory DNA elements located at the promoter or at more distant regions such as enhancers, silencers or insulators (Smale and Kadonaga, 2003). In addition to genetic factors, the increase of genome-wide studies has revealed a growing number of epigenetic factors now known to regulate gene transcription. Most of these factors are proteins but a number of studies have also identified ncRNAs as possible regulators of gene transcription (Goodrich and Kugel, 2006).

Ultimately, the spatial organization of the genome is also important for efficient transcription. For instance, the process of transcription is spatially compartmentalized in transcription factories (Iborra et al., 1996), thus efficiency of transcription may be enhanced if several actively transcribed genes associate with shared transcription and processing sites (Sutherland and Bickmore, 2009).

1.1.1 *Transcription initiation*

The first step of eukaryotic gene expression is the recruitment of RNAPII and the necessary cofactors to a gene promoter. This step is known as the preinitiation step and involves the formation of a preinitiation complex (PIC) composed of RNAPII and general transcription factors (GTFs). The PIC helps to position RNAPII transcription start site (TSS) of genes (Maston et al., 2006). Typically the PIC is composed of six

general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH but various combinations of core promoter motifs target the assembly of distinct PICs (Juven-Gershon et al., 2008). *In vitro*, the assembly of a PIC on the core promoter is sufficient to initiate transcription, but *in vivo* activators and co-regulators are needed to accurately initiate transcription. An important complex recruited at the preinitiation step *in vivo* is the Mediator, a large complex composed of more than 20 distinct subunits highly conserved from yeast to humans (Conaway et al., 2005). The Mediator is known to bind to the carboxy-terminal domain (CTD) of RNAPII, acting to stabilize the formation of the PIC (Bjorklund and Gustafsson, 2005). Interestingly, other functions of Mediator might include the promotion and/or stabilization enhancer-promoter interactions, with the help of cohesin, as reported in a recent study (Kagey et al., 2010).

In addition to the "classic" GTFs and the Mediator complex, activators are also required for accurate transcription. Activators bind to specific DNA motifs within promoters or enhancers and, by direct interactions with one or more components of the transcriptional machinery, can promote PIC assembly or other steps of the transcription process subsequent to PIC formation, such as initiation, elongation, or re-initiation (Lee and Young, 2000). Activators have also been proposed to function by the recruitment of co-regulators.

Co-regulators can either enhance (co-activators) or repress (co-repressors) transcription, typically by chromatin-modifying activities which act to make the DNA more or less accessible to the transcription machinery (Kadonaga, 2004). Examples of genes regulated in this manner include those that require activators to interact with Swi/Snf (SWI/SNF in humans) remodeling complexes to enhance transcription from nucleosome covered promoters (Peterson and Workman, 2000). After RNAPII is recruited, local opening of the double-stranded DNA (also known as promoter melting) takes place to allow for RNAPII to initiate transcription. Transcription factor II H (TFIIH), a large multisubunit complex, is needed at this step. TFIIH contains ATP-dependent helicases that act on the DNA template to promote its opening and allow RNAPII to initiate transcription. TFIIH is also required to promote RNAPII release from the promoter-bound factors. This step, known as

"promoter escape", depends on the phosphorylation of serine 5 residues of the CTD (by Cdk7 subunit of TFIIH; subsection 1.3.1) and on the Mediator complex (Hirose and Ohkuma, 2007). Due to the instability of the RNAPII complex, initiating RNAPII can abort transcription, usually after producing ~10 nucleotides (nt), a phenomena known as "abortive initiation".

1.1.2 *Transcription Elongation*

Initiating RNAPII escapes from the promoter and is then able to generate short transcripts; however, elongation can be halted after RNAPII transcribes ~25-50nt by negative factors which include the 5,6-dichloro-1-b-D-ribofuranoxylbenzimidazole sensivitiy-inducing factor (DSIF) and negative elongation factor (NELF) protein complexes (Fuda et al., 2009). Elongation requires the recruitment of chromatin remodeling factors and the positive transcription elongation factor b (P-TEFb). P-TEFb contains a kinase activity (Cdk9) that phosphorylates RNAPII CTD on its serine 2 (S2) residues (Cheng and Price, 2007; subsection 1.3.2).

Although much emphasis has been given to the regulation of gene transcription at the step of RNAPII recruitment (Nevado et al., 1999), a growing number of studies indicates that many inactive genes recruit RNAPII to their 5′ ends in both mammals (Guenther et al., 2007, Kim et al., 2005, Rahl et al., 2010) and *Drosophila* (Muse et al., 2007, Zeitlinger et al., 2007), suggesting that many genes in the genome are regulated at steps following the recruitment of RNAPII to a promoter.

1.1.3 *Transcription termination*

When RNAPII reaches the poly-adenylation signal at the 3′ end of genes, transcription termination takes place. The process of transcription termination involves the correct poly-adenylation of the newly formed mRNA and its release, the digestion of downstream transcript by exonucleases, and the dissociation of the RNAPII transcription complex from the chromatin (Richard and Manley, 2009). Transcription termination has also been shown to be important to prevent aberrant transcription from downstream genes (Greger et al., 2000). RNAPII is then targeted

by CTD phosphatases, such as Fcp1, and once in its hypophosphorylated form can then re-initiate a new cycle of transcription (Cho et al., 2001).

1.1.4 *Core promoter elements and gene regulation*

Core promoter elements have a key role in contributing to gene regulation control, as they constitute the docking site where PIC assembles, and define the position of the TSSs and the direction of transcription itself (Smale and Kadonaga, 2003). A number of DNA elements have been found at Metazoan core promoters, such as TATA box, Initiator element (Inr), Downstream Promoter Element (DPE), upstream TFIIB Recognition Element (BREu), Motif Ten Element (MTE) and Downstream Core Element (DCE). Combinations of different DNA motifs at promoters specify the binding of different GTFs and different responses to enhancers (Juven-Gershon et al., 2008).

Interestingly, about 60% of the human genes overlap a CpG island. CpG islands are DNA sequences typically 500 bp to 2 kb long with high frequency of CpG dinucleotide compared to the bulk of DNA in the mammalian (Illingworth and Bird, 2009). CpG island promoters (CGI promoters) are normally associated with constitutively expressed housekeeping functions or other highly regulated functions, such as developmental regulation (Illingworth and Bird, 2009). CpG islands form a chromatin environment that is generally more permissive to transcription (Weber et al., 2007), possibly due to low nucleosome occupancy at these promoters (Ramirez-Carrozzi et al., 2009).

Sequencing of transcript 5′ ends, namely through the use of CAGE and DeepCAGE techniques (Valen et al., 2009, Carninci et al., 2006), has expanded our knowledge on the characterisation of mammalian promoters in different cell types and revealed the existence of many genes with several potential TSS within the same promoter. This "broad" type of promoters are characteristic of CGI promoters, while "sharp" promoters, which initiate from a single TSS, are generally associated with TATA box, Inr or DPE binding motifs (Carninci et al., 2006, Juven-Gershon and Kadonaga, 2010). The variety of core promoter elements and their combinations and shapes suggests a key role of promoters in the specificity of gene expression control.

1.1.5 *Functions of gene clustering*

Apart from DNA elements, such as core promoters and enhancers that influence recruitment and the assembly of the transcription machinery, there is another layer of information embedded within the DNA sequence which contributes to gene regulation. Gene density and the genomic context of a given gene or a cluster of genes provide an important level of gene expression control (De and Babu, 2010).

At a higher level, the organization of the chromosomes within the nucleus is nonrandom; chromosomes occupy discrete positions in the nucleus known as "chromosome territories" (Meaburn and Misteli, 2007, Cremer et al., 2006). At the chromatin-scale, areas of condensed and de-condensed chromatin can be distinguished within chromosomes. Condensed chromatin is generally associated with H3K9me3, H3K27me3 and repressive proteins such as HP1, while de-condensed chromatin (termed euchromatin) is associated with H3K9ac, H3K4me3 and H3K36me3. There are also many examples of physical interactions between different chromatin segments resulting in chromosomal loops. One of the most studied examples of chromatin folding comes from the *ß-globin* locus, where it has been shown by DNA FISH and RNA FISH and chromosome conformation capture methodologies (such as 3C, 4C) that active *ß-globin* genes interact with other active genes located several Mb away (such as *Uros, Eraf*, and *Kcnq1*; Osborne et al., 2004, Mitchell and Fraser, 2008, Simonis et al., Dekker, 2003). Other examples of chromatin movements relate to silencing. For instance, upon differentiation of ES cells to neuronal progenitor cells (NPCs), the pluripotency genes *Oct4* and *Nanog* are positioned at the nuclear lamina, a compartment in the periphery of the nucleus which is inhibitory to transcription, while silent neuronal-lineage specific genes detach from the lamina and become active (Peric-Hupkes et al., 2010).

A. Genomic context

The genomic context in which the genes are embedded also seems to influence efficient and reliable gene expression (Madan Babu et al., 2008). For instance, transfection of a human transcriptional unit with the required promoter and enhancer elements resulted in gene silencing if the transgene unit, lacking its LCR, is

integrated within heterochromatic regions (Festenstein et al. 1996). Additionally, a more or less compact genomic context could influence chromatin movements. An example of such influence comes from the the increased re-organization of the α-globin locus in erythroid cells in humans compared to mouse, possibly due to its genomic context which, in humans, consists of greater gene density and a more open chromatin structure (Brown et al., 2008).

B. Clustering of genes in the linear DNA sequence

If the chromatin environment does contribute to accurate gene expression, then this could create an evolutionary pressure in the way that genes are clustered in the chromosome (Batada and Hurst, 2007, Sproul et al., 2005, De et al., 2009, Janga et al., 2008). In fact, genomic regions of highest gene density are also the ones that contain most highly and broadly expressed genes in the genome (Versteeg et al., 2003). Broadly expressed genes are also generally GC rich and have short introns (Versteeg et al., 2003). Gene clustering could also reflect coordinated gene expression. For instance, genes that are coordinately regulated during development, or that are expressed in a specific tissue, are often found in clusters (Sproul et al., 2005). Moreover, in multicellular eukaryotes, co-expressed genes form larger clusters than in yeast and worm (Hurst et al., 2004), thus reflecting an evolutionary pressure for gene clustering that could relate to mechanisms of gene expression control.

1.1.6 *Epigenetic gene regulation*

Nucleosomes form the basic unit of chromatin in eukaryotes, consisting of ~147 base pairs of DNA wrapped around histone octamers (normally consisting of two each of H2A, H2B, H3, and H4). Histone octamers are separated by linker DNA which is associated with a fifth histone, H1. Post-translational histone modifications, driven by enzymatic complexes that add or remove covalent groups (e.g., acetyl groups and methyl groups) to histone tails, are an important epigenetic mechanism of gene regulation (Kouzarides, 2007). Several histone modifications have been described, including methylation, phosphorylation, acetylation, ubiquitination and ADP ribosylation. Histone modifications can impact on chromatin compaction, thus

affecting the accessibility of the promoters to the transcription machinery. They also exert a more direct effect on the transcriptional status of genes by being directly involved in the recruitment of specific binding factors such as chromatin remodelers. Chromatin remodelers are ATP-dependent protein complexes that non-covalently modify chromatin structure. Several mechanisms by which chromatin remodelers act have been elucidated, including the capability to influence nucleosome positioning (Li et al., 2007, Gutierrez et al., 2007). SWI/SNF remodeling complex, have been shown to be required for the regulation of pluripotency and differentiation, as disruption of several members of these complex is embryonic lethal (reviewed in Keenen and de la Serna, 2009). In addition to the described histones, "histone variants" (such as H2AZ, H2AX and H3.3) contribute to nucleosome diversity and comprise another layer of epigenetic regulation (Henikoff et al., 2004). Recent work showed that histone variant H2AZ and the Polycomb subunit Suz12 co-occupy the same promoters in ES cells and demonstrated that H2AZ is necessary for ES cell differentiation (Creyghton et al., 2008).

The availability of antibodies capable of distinguishing between different lysine positions and methylation states allowed identification of methylation patterns by ChIP-Chip and ChIP-Seq methodologies on a genome-wide scale. While histone acetylation is associated with gene activation and open chromatin conformation, methylation correlates with activating or repressing chromatin depending on the residue of the histone tail being modified. E.g., H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 36 trimethylation (H3K36me3) and monomethylation on H3K4, H3K9, H3K27, H4K20, and H2BK5 correlate with gene activation, while trimethylations on H3K9 and H3K27 are linked to repression (Kouzarides, 2007). Studies in human and mouse ES cells, neural progenitor cells (NPCs), fibroblasts, hepatocytes and T cells have revealed patterns of histone methylation at promoters (Barski et al., 2007, Guenther et al., 2007, Mikkelsen et al., 2007, Pan et al., 2007), enhancers and insulators (Barski et al., 2007, Heintzman et al., 2007), and gene body regions (Kolasinska-Zwierz et al., 2009, Cui et al., 2009). In this thesis, I focus on discussing H3K4me3 and H3K36me3 which are enriched at the promoter and gene body regions, respectively, of actively transcribed genes, and the Polycomb-instigated

histone modifications H3K27me3 and H2Aub1 which are largely correlated with gene repression.

Methylation of DNA cytosine residues is an important epigenetic silencing mechanism associated with silenced chromatin states. DNA methylation occurs at the DNA cytosine base, typically in a CpG dinucleotide context and is mediated by DNA methyltransferases (DNMTases, Dnmt1, Dnmt3a, Dnmt3b). In mammals, CpG-poor promoters are more likely to be silenced by DNA-methylation mediated events, whereas CpG-rich sequences are largely unmethylated (Fouse et al., 2008, Meissner et al., 2008, Mohn and Schubeler, 2009). DNA methylation is crucial in downregulating genes associated with late-differentiation (Fouse et al., 2008) and has also been shown to be involved in development by homozygous (Li et al., 1992) deletion of Dnmt1, which results in lethality during early development in mice. Despite its importance for normal development, it has been proposed that DNA methylation is not the major mechanism of transcriptional repression in ES cells because promoter methylation might be incompatible with the ability of stem cells to activate a wide range of lineage-specific genes during differentiation (Reik, 2007). In agreement, Polycomb-mediated gene repression, a more dynamic form of epigenetic silencing, is observed at many developmental regulators in human and mouse ES cells (Ku et al., 2008, Boyer et al., 2006, Lee et al., 2006). In this introduction, I discuss Polycomb-mediated gene repression in more detail in section 1.5.

1.1.7 *Gene regulation in ES cells*

Derived from the inner cell mass of the mammalian blastocyst, ES cells have the ability to differentiate into all somatic cell lineages of an organism, the phenomenon of pluripotency (Evans et al., 1981, Martin, 1981). In addition, if maintained under the appropriate conditions *in vitro*, namely in the presence of recombinant leukemia inhibitory factor (LIF) and fetal bovine serum (FBS; Williams et al., 1988), ES cells display unlimited ability for self-renewal. Understanding the mechanisms of gene regulation in pluripotent ES cells and the molecular switches that regulate gene expression upon differentiation will allow us to better utilize their potential in therapeutic applications. Key 'pluripotency' factors Oct4, Nanog and Sox2 constitute

the core regulatory network that governs ES cell pluripotency. These core transcription factors form a tight regulatory network, in which they autoregulate each other and synergistically regulate downstream genes, controlling both the normal ES cell maintenance functions, as well as the differentiation pathways (Silva and Smith, 2008). Aside from transcription factors, chromatin-based regulatory mechanisms are important to modulate gene expression in ES cells.

ES cells are characterized by a global chromatin structure that is generally decondensed, which is thought to be functionally relevant for pluripotency (Mattout and Meshorer, 2010). Knockdown of Chd1, a chromatin remodeler, results in increased chromatin condensation and loss of potential to differentiate (Gaspar-Maia et al., 2009), thus demonstrating the importance of generally "open" ES chromatin structure. In ES cells, heterochromatin is more dispersed microscopically than in NPCs (Meshorer and Misteli, 2006), with major structural proteins, such as histone protein HP1, being more loosely bound to DNA in undifferentiated ES cells compared to NPCs (Meshorer and Misteli, 2006). Positioning of some gene loci and chromosome territories have been shown to change markedly upon ES cell differentiation (Chambeyron and Bickmore, 2004, Williams et al., 2006). ES cells also exhibit a higher overall rate of transcription (Efroni et al., 2008, Araki et al., 2006, Turner, 2008).

Silenced developmental genes are bound by Polycomb proteins and enriched for H3K27me3 in ES cells (Boyer et al., 2006, Lee et al., 2006). PRC-target genes also exhibit features of active chromatin, including H3K9ac and H3K4me3, which is thought to prime genes for subsequent activation (Azuara et al., 2006, Bernstein et al., 2006). Genes with concomitant H3K27 and H3K4 trimethylation have been termed 'bivalent genes', as they display histone modifications that are related to opposite effects on transcription. Polycomb proteins (section 1.5) and their interplay with RNAPII complexes (section 1.3) in ES cells is the main focus of this thesis.

1.2 RNA POLYMERASE II AND THE C-TERMINAL DOMAIN

In eukaryotes, the family of nuclear RNA polymerases consists of three RNA polymerases (RNAPI, RNAPII and RNAPIII). RNA polymerase I and III mainly

transcribe structural RNAs. RNAPI transcribes a single transcript, the precursor rRNA. RNAPIII transcribes transfer RNAs and 5S rRNAs and has more recently been shown to transcribe small RNAs that can regulate expression of protein coding genes (Dieci et al., 2007). In contrast, RNAPII is responsible for transcription of all protein-coding genes, and a variety of structural and regulatory noncoding RNAs (ncRNAs).

RNAPII is a large multisubunit complex, consisting of 12 subunits (RPB1 to 12). Its largest subunit, RPB1, contains a carboxy-terminal domain (CTD) consisting of multiple tandem heptapeptide repeats of consensus sequence N-Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7-C, with 52 copies in mammals (Corden et al., 1985). CTD length increases with the organismal complexity, having 44 repeats in *Drosophila* and 26 to 27 repeats in yeast (Chapman et al., 2008). The crucial CTD function remained unknown until it was discovered, in late 80s, that deletion of CTD to less then 10 repeats was lethal in yeast (Nonet et al., 1987). The CTD is not required for the catalytic activity of RNAPII *in vitro*, but essential for transcription *in vivo* (Meininghaus et al., 2000). The CTD has a highly flexible, mobile structure, which is generally more compact in the unphosphorylated state, and adopts a more extended form upon extensive phosphorylation possibly due to charge repulsions (Meinhart et al., 2005). Currently, it is known that recruitment of appropriate factors at different stages of the transcription cycle is achieved through changing patterns of post-translational modification of the CTD repeats creating a readable 'code'. The CTD thus helps integrate the synthesis of RNA with local chromatin remodeling and co-transcriptional RNA processing (Hirose and Ohkuma, 2007). Moreover, CTD modification patterns were also observed to change during cell cycle, virus infection and DNA damage (Jeong et al., 2005, Bregman et al., 2000). Many different CTD modifications have been identified, including phosphorylation, isomerisation, glycosylation (Sobennikova et al., 2007) and, more recently, methylation (Sims et al., 2011).

1.3 INTEGRATION OF THE CTD CODE WITH REGULATORY PROCESSES

1.3.1 *CTD serine* 5 *phosphorylation and transcription initiation*

At transcription initiation, CTD repeats are phosphorylated at the S5 residue by the kinase subunit of TFIIH (Lee and Young, 2000, Liu et al., 2004). TFIIH is a large protein complex of 10 subunits in mammals and its kinase activity is sustained by the Cdk7/Cyclin H pair. Mammalian Cdk8/cyclin C, a component of the Mediator complex, is another enzyme capable of phosphorylating S5 residues (Liao et al., 1995, Rickert et al., 1999) but if this phosphorylation occurs prior to the formation of the preinitiation complex, it has a negative effect on transcription (Hengartner et al., 1998). Erk1 and Erk2 kinases are also known to phosphorylate S5 during stress responses, thereby thought to be associated with the regulation of specific genes in response to extreme conditions (Sobennikova et al., 2007).

At initiation of transcription, S5 phosphorylation promotes RNAPII clearance from the promoter (Liu et al., 2004, Max et al., 2007). S5P is also important for the recruitment of the RNA capping enzymes (Ho and Shuman, 1999, Rodriguez et al., 2000), thus integrating transcription with RNA processing (Figure 1.1). Several studies have shown that the initiating form of RNAPII (S5P but not S2P) can recruit Set1 protein in yeast (Ng et al., 2003) and Mll1 in mouse (Milne et al., 2005), both proteins are major H3K4 methylases, indicating that the presence of H3K4me3 at gene promoters may provide a molecular memory of recent transcriptional activity. However, RNAPII-independent recruitment of H3K4me3 HMTs has also been reported (Edmunds et al., 2007), thereby raising the possibility that RNAPII recruitment and H3K4me3 deposition are not necessarily sequential, but that reciprocal interactions may also occur. In any case, independently of the order of events (RNAPII recruitment or H3K4me3 deposition first), it is evident that H3K4me3 has a positive effect on transcription. For instance, Taf3 (a TFIID component) has been shown to bind to H3K4me3 (Vermeulen et al., 2007). Interactions between H3K4me3 and members of the PIC are thought to facilitate stable assembly of the PIC at the onset of transcription (Dambacher et al., 2010). Moreover, H3K4me3 deposition was shown to act as a binding platform for chromatin remodelers, such as

BPTF (a nucleosome remodeling factor (NURF) subunit) and Chd1 (Flanagan et al., 2005, Wysocka et al., 2006) which act to slide nucleosomes thus creating an "open" chromatin structure more accessible to transcription initiation.

The current model (Figure 1.2) of the transcription cycle, sees S5P phosphorylation occuring early in the transcription cycle and gradually decreasing towards the 3′ end of genes (Komarnitsky et al., 2000, Tietjen et al., 2010, Figure 1.1). S5P removal is mediated by phosphatases, such as Rtr1 and Scp1, which are recruited shortly after RNAPII release from the promoter (Mosley et al., 2009, Zhang et al., 2006b).

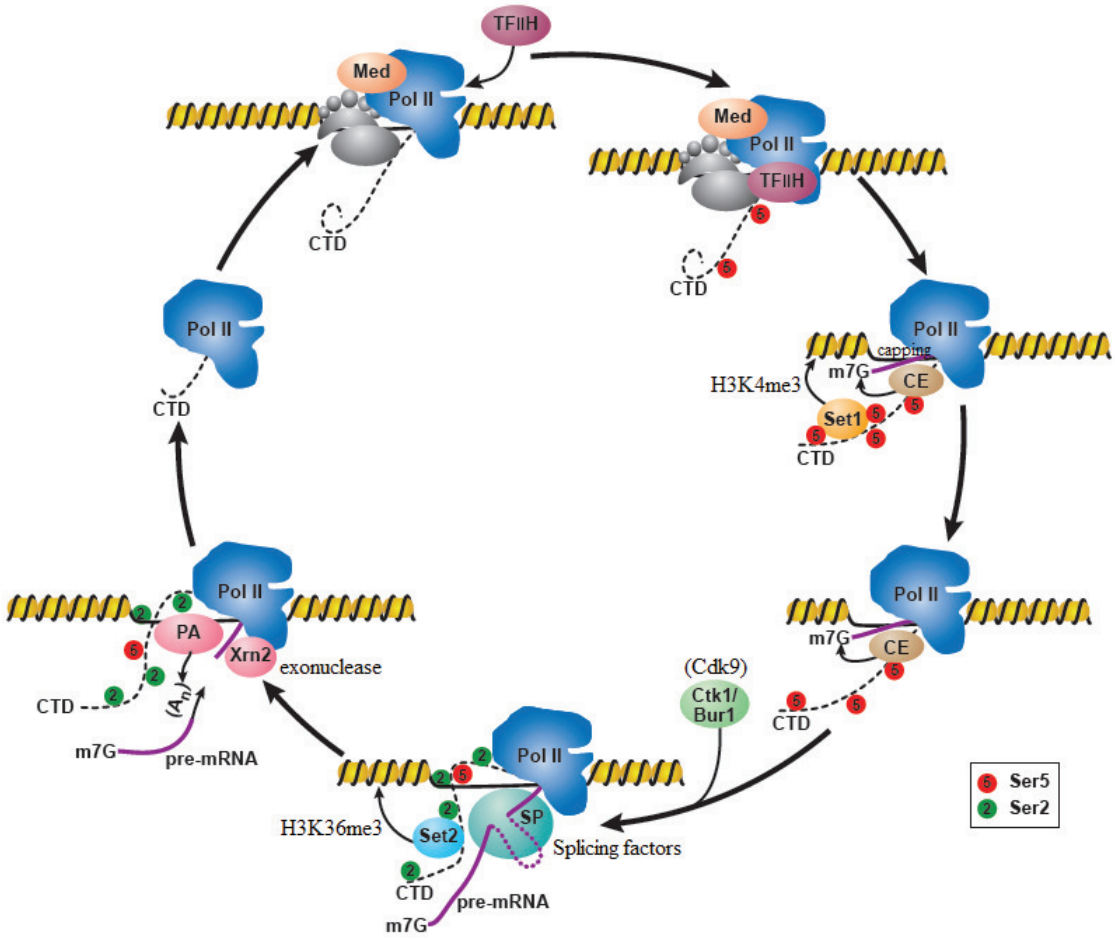


Figure 1.1: Schematic representation of the transcriptional cycle. RNAPII is recruited at gene promoters together with Mediator (Med) and TFIIH. During transcription initiation, CTD is phosphorylated on serine 5 which enhances the binding of capping enzymes (CE) and Set1, the HMT responsible for H3K4me3. During elongation, serine 2 residue is phosphorylated by Ctk1 and Bur1 (Cdk9 in mammals). S2 recruits several elongation factors, including co-transcriptional splicing factors (SP) and also Set2 which trimethylates H3K36. During transcription termination, the newly formed mRNA is polyadenylated and exonucleases (Xrn2) digest the downstream transcript. Adapted from Tietjen et al. [2010].

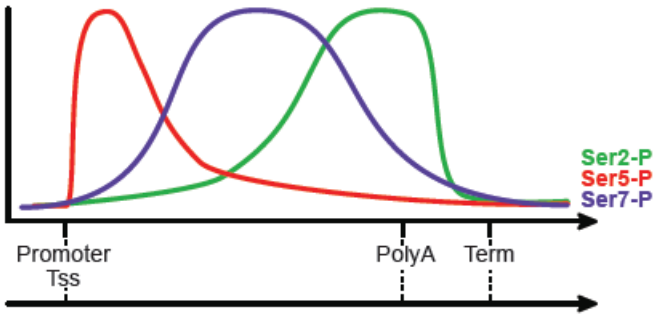


Figure 1.2: The current model of CTD phosphorylation patterns. Initially an enrichment of S5P peaks at the promoter and is rapidly dephosphorylated shortly after transcription initiation. S2P and S7P are enriched through gene bodies. Adapted from Tietjen et al. [2010].

1.3.2 *CTD serine 2 phosphorylation and transcription elongation*

Serine 2 phosphorylation (S2P) is mediated by the Cdk9 kinase subunit of P-TEFb complex in mammals (Marshall et al., 1996), and Ctk1 and Bur1 in yeast (Wood and Shilatifard, 2006). At actively transcribed genes, S2P increases throughout the gene body towards the 3′ end of genes (Komarnitsky et al., 2000, Tietjen et al., 2010, Figure 1.2).

S2P is known to be important for efficient elongation (Figure 1.1), integrating transcription with mRNA co-transcriptional processing. S2P marks are recognized by splicing factors (Kornblihtt et al., 2004) and 3′ end processing proteins (Proudfoot et al., 2002, Ahn et al., 2004). In addition, Set2 enzyme, the HMT capable of trimethylating H3K36, was shown to be recruited by RNAPII-S2P, thereby integrating transcription with chromatin modifications (Krogan et al., 2003, Xiao et al., 2003, Li et al., 2003). In fact, the H3K36me3 HMT, Set2, was shown to preferentially recognise doubly marked CTD, S5P and S2P, *in vitro*, rather then S2P alone (Kizer et al., 2005), thus contributing to the idea that both S5P and S2P are present during transcription elongation.

H3K36me3 modified chromatin is thought to be important for mRNA processing regulation. For instance, it was recently shown that H3K36me3 is particularly enriched at the exons of actively transcribed genes, suggesting a role of this modification in splicing regulation (Kolasinska-Zwierz et al., 2009). Other studies

questioned the role of H3K36me3 in splicing (Tilgner et al., 2009) arguing that the enrichment of H3K36me3 at exons and splice sites simply reflects increased nucleosome density. Another study, however, by Luco et al., 2010 showed that by modulating the levels of H3K4me2, H3K9ac, H3K27ac and H3K36me3, different splice isoforms of the human *FGFR2* gene were transcribed, thus adding further evidence to the role of H3K36me3 in splicing and specially in alternative splice choices. A role of H3K36me3 in selection of 3′ end polyadenylation sites has also been suggested (Spies et al., 2009). Furthermore, H3K36me3 was shown to recruit Mrg15 (Zhang et al., 2006a) which in turn depletes intragenic H3K4me3 (Hayakawa et al., 2007) and recruits histones deacetylases (Yochum and Ayer, 2002). It is thought that intragenic deacetylation and H3K4me3 depletion are important to prevent aberrant transcription initiation from within the gene bodies (Dambacher et al., 2010). Taken together, these examples demonstrate how different proteins and histone marks cooperate to allow robust and efficient gene transcription. The complexity of this dynamic interactions demonstrate how transcription can be regulated at multiple levels, for instance by altering chromatin states or blocking specific mRNA processing factors, which thereby generates sophisticated modes of gene control, greatly increasing gene regulatory potential.

1.3.3 *CTD serine 7 phosphorylation*

Serine 7 residue of the CTD is part of the canonical heptapeptide CTD repeat (N-Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7-C), conserved from yeast to mammals. The CTD has degenerated sequences typically after the 21^{st} repeat and predominantly on the serine 7 position, in humans. Non-canonical repeats are thought to add complexity and functionality to the CTD (Chapman et al., 2007) but their role is still unclear as their replacement with random repeats did not affect cell viability (Chapman et al., 2005).

S7, like the other serine residues of the CTD, S2 and S5, can be phosphorylated, but its functions during transcription have only recently start to be revealed. S7 phosphorylation was found to occur at protein-coding genes in yeast and humans (Chapman et al., 2007, Kim et al., 2010a, Tietjen et al., 2010, Mayer et al., 2010) and at

snRNA genes in humans (Egloff et al., 2007). S7, like S5, is phosphorylated by Cdk7, the catalytic subunit of TFIIH, both *in vivo* and *in vitro* (Akhtar et al., 2009, Glover-Cutter et al., 2009), and to be dependent on the Mediator complex (Boeing et al., 2010). P-TEFb kinase subunits, Cdk9 (mammals) and Bur1 (yest), were also shown to be able to phosphorylate S7 (Glover-Cutter et al., 2009, Tietjen et al., 2010).

Currently, it is known that S7P is required for Integrator binding and 3′ processing of snRNAs (Egloff et al., 2010). The integrator is a multisubunit complex responsible for snRNA 3′ end recognition and processing (Baillat et al. [2005]). In fact, phosphorylation of both S2P and S7P is necessary for Integrator binding at snRNA genes (Egloff et al., 2010). In yeast protein-coding genes, S7P was found at the gene bodies (Kim et al., 2010a, Mayer et al., 2010, Tietjen et al., 2010; Figure 1.2), and particularly enriched close to 5′ splice site (Kim et al., 2010a), suggesting a role of this mark in co-transcriptional splicing (Kim et al., 2010a, Tietjen et al., 2010).

1.3.4 *Other CTD modifications*

A CTD repeat can be modified at other different positions (reviewed in Sobennikova et al. 2007). Proline residues (P3 and P6) can exist in either *cis* or *trans*, while the other five amino acid residues can be phosphorylated and glucosylated. In higher eukaryotes additional residues appear at degenerate CTD repeats, namely lysines and arginine. They can potentially be methylated or acetylated, thus increasing the combinatorial potential of the "CTD code". One example is the newly identified arginine methylation in the 31^{st} repeat of the human CTD (Sims et al., 2011). In many cases, the function of these modifications, whether it affects transcription and pre-mRNA processing, remains unclear (summarised in Table 1.1).

Table 1.1: CTD modifications and respective modifying enzymes and their effect on transcription.

CTD repeat and modification	Function in transcription	Enzymes
Serine 5 phosphoprylation (S5P)	Initiation, recruitment of capping enzymes, recruitment of Set1 (H3K4me3 HMT)	Cdk7 (TFIIH)
Serine 2 phosphorylation (S2P)	Elongation, regulation of splicing and 3′ processing, recruitment of Set2 (H3K36me3 HMT)	Cdk9 (P-TEFb)
Serine 7 phosphorylation (S7P)	Regulation of splicing, snRNA 3′ end processing	Cdk7 (TFIIH) Cdk9 (P-TEFb)
Proline isomerization	Changes in CTD structure. Thought to be important in initiation (Krishnamurthy et al., 2009; Xu and Manley, 2007) and mRNA processing (Hani et al., 1999)	Pin1
Tyrosine 1 phosphorylation (Y1P)	Thought to be important for efficient elongation of specific genes (Baskaran et al., 1999)	c-Abl
Threonine 4 phosphorylation (T4P)	Not characterised	Not characterised
Serine glycosylation Threoninne glycosylation	Not characterised	Not characterised O-GlcNAc transferase
Arginine methylation	Thought to repress the expression of snRNAs and snoRNAs (Sims et al., 2011)	CARM1

1.4 TRANSCRIPTIONAL PRIMING AND RNAPII PAUSING

1.4.1 *RNAPII pausing*

RNAPII "promoter-proximal" pausing is defined as RNAPII accumulation downstream of TSS of genes and is associated with 5′ capped transcripts which are typically 25-50 nucleotides long (e.g. Nechaev et al., 2010), and stable enough to be detected by permanganate protection assays (Rougvie and Lis, 1988), and/or the genome-wide run-on assay Gro-Seq (Core et al., 2008). In addition, paused RNAPII is detectable by ChIP using 8WG16 (as in Kim et al., 2005, Guenther et al., 2007, Boehm

et al., 2003), an antibody that has preference for unphosphorylated S2 residues. RNAPII pausing was first described at heat shock genes in *Drosophila* (Rougvie and Lis, 1988). At these genes, heat shock acts as an activating signal which induces the rapid release of RNAPII into transcription. Several subsequent studies, particularly in *Drosophila*, led to a detailed understanding of RNAPII pausing as a mechanism of gene control (reviewed in Fuda et al., 2009, Levine, 2011). Initially, this mode of transcription regulation was thought to be specifically used by heat shock genes or other stress responsive genes that were quickly activated in response to extreme conditions. However, with the advent of genome-wide approaches, the understanding of RNAPII has evolved, and it became clear that it is a common regulatory event, not necessarily limited to stress response, as it is also observed at many genes including inactive and actively transcribed genes, in *Drosophila* (Muse et al., 2007, Zeitlinger et al., 2007, Gilchrist et al., 2010, Nechaev et al., 2010), mouse (Rahl et al., 2010) and human (Guenther et al., 2007, Core et al., 2008, Min et al., 2011). These observations support the view of pausing as an early "checkpoint" which regulates RNAPII release into productive elongation.

Several mechanisms by which RNAPII pausing can be advantageous have been purposed. Gilchrist et al., 2010 suggested that pausing is crucial to block nucleosome assembly at highly regulated genes whose "default" chromatin state is prone to nucleosome assembly. Previous studies have shown that RNAPII pausing affects nucleosome positioning in humans (Schones et al., 2008). In addition, RNAPII pausing was suggested to facilitate mRNA capping (Pei et al., 2003) and promoter-insulator interactions (Chopra et al., 2009a). In mammals, most paused promoters are associated with upstream anti-sense transcription of unstable RNAs (Core et al., 2008, Seila et al., 2008). It has been proposed that RNAPII pausing may constitute a mechanism of regulating transcription from bidirectional promoters (Rahl et al., 2010).

1.4.2 *Control of RNAPII pausing*

The release of RNAPII from promoter-proximal regions into the main body of the gene is blocked by pausing factors which include the multisubunit complexes DSIF

and NELF (Yamaguchi et al., 1999, Wu et al., 2003). Recent studies suggest that the Spt5 subunit of DSIF binds to nascent RNA as they are transcribed from RNAPII (Missra and Gilmour, 2010). NELF is recruited by DSIF, and the two complexes prevent RNAPII progression beyond the pause site. NELF was found to be enriched at active genes in *Drosophila* (Lee et al., 2008), consisting with a model of pausing as widespread event that regulates transcription at active genes.

Recruitment of P-TEFb causes the dissociation of NELF and serine 2 phosphorylation, resulting in the release of RNAPII into productive elongation through the gene body (Keogh et al., 2003, Boehm et al., 2003). Dissociation of NELF from the paused RNAPII complex is triggered by its phosphorylation, mediated by a kinase Cdk9 of P-TEFb (Fujinaga et al., 2004). Cdk9 also phosphorylates DSIF resulting in its activation as a positive elongation factor (Yamada et al., 2006).

1.4.3 *RNAPII at silent Polycomb repressed genes*

In this thesis, I will show that RNAPII binding to silent genes as a mechanism of gene priming is especially relevant in the context of ES cells, which have to maintain their genome in a potentiated state, able to respond to differentiation cues. Polycomb group of proteins (PcG) are a major epigenetic mechanism of gene regulation in ES cells (Lund and van Lohuizen, 2004). PcG proteins were initially thought to prevent RNAPII recruitment to the promoters of its target genes (Dellino et al., 2004); but, interestingly, several studies have demonstrated that RNAPII is recruited at PRC-repressed genes (Brookes and Pombo, 2009). PRC-repression and RNAPII priming in ES cells form the main focus of this thesis.

1.5 THE POLYCOMB REPRESSIVE COMPLEXES

1.5.1 *Structure and function of Polycomb complexes*

Polycomb group (PcG) proteins were originally described in *Drosophila melanogaster* as repressors of homeotic genes such as *Hox* genes, acting in opposition to Trithorax proteins. In mammals, two distinct Polycomb remodeling complexes mediate transcriptional repression: Polycomb-repressive complex 1 (PRC1) and

Polycomb-repressive complex 2 (PRC2; Lund and van Lohuizen, 2004). PRC2 contains as core components, embryonic ectoderm development (Eed), suppressor of zeste 12 (Suz12) and enhancer of zeste homologue 2 (Ezh2), which catalyzes di and tri-methylation of histone H3 at lysine-27 (H3K27me2 and H3K27me3). PRC1 is composed of approximately 10 subunits, including Ring1A and Ring1B, Bmi1 and Cbx8, and mediates repression by Ring1B subunit which acts as a ubiquitin E3 ligase specific for histone H2A lysine 119 (Wang et al., 2004). Ring1A and Bmi1 positively regulate Ring1B-mediated H2A ubiquitination (Cao et al. 2005). Ring1A was also shown to have E3 ligase activity capable of ubiquitinating H2A *in vitro* (Buchwald et al. 2006). *In vivo*, however, Ring1B has a predominant role as Ring1B depletion results in embryonic lethality, while Ring1A knockout mice are viable with minor skeletal transformations (del Mar Lorente et al. 2000, Voncken et al. 2003).

The H3K27me3 modification catalyzed by PRC2 is thought to serve as binding platform for PRC1 (Figure 1.3), which further enhances PRC2 repression (Cao et al., 2002). However PRC1 recruitment independent PRC2 was reported in particular during X inactivation in differentiating ES cells (Schoeftner et al., 2006). Moreover, genes bound solely by PRC2 have also been found in mouse ES cells (Ku et al. 2008). Taken together, these results suggest that PRC1 recruitment can be regulated at different stages, and that PRC2 binding can be uncoupled from PRC1 recruitment.

A novel component of PRC2, Jarid2, has recently been implicated in ES cell differentiation. Jarid2 contains a Jumonji domain (JmjC), enzymatically inactive. Jarid2 knockdown impairs ES cell differentiation, without preventing self-renewal (reviewed in Landeira and Fisher, 2010). Jarid2 has been implicated in the recruitment of PRC2 and PRC1, and in the establishment and/or maintainence of RNAPII-S5P in bivalent genes in mouse ES cells (Landeira et al. 2010).

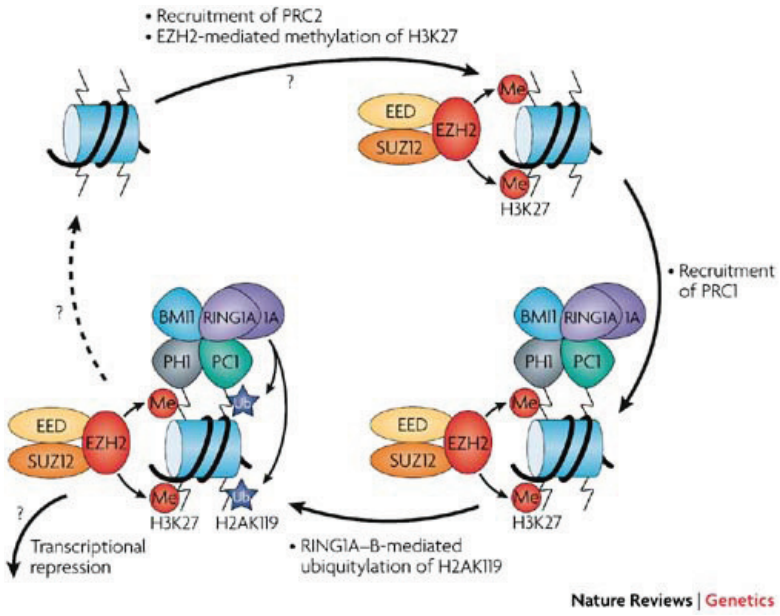


Figure 1.3: Polycomb repressive complexes. PRC2 contains as core components Ezh2, Eed and Suz12. Ezh2 mehyltransferase methylates lysine 27 of histone H3 (H3K27me). This modification recruits PRC1. PRC1 contains proteins such as Bmi1, Ring1B and Ring1A. Ring1B acts as a ubiquitin E3 ligase which monoubiquitinates histone H2A on lysine 119 (H2AK119ub1). Image from Spivakov and Fisher [2007].

1.5.2 *Polycomb-mediated gene silencing*

The importance of Polycomb Group proteins in developmental regulation is well established. Genome-wide mapping of PRC1 and PRC2 components in mouse (Boyer et al., 2006, Ku et al., 2008) and human (Lee et al., 2006) ES cells revealed that many PRC-targets include genes encoding homeodomain transcription factors such as *Hox*, *Pax*, *Irx* and *Six* as well as other developmental regulators such as *Fox* and *Sox* that are preferentially expressed upon differentiation (Boyer et al., 2006). PcG proteins are thought to play a crucial role in maintaining ES cells' pluripotent state by binding to the promoters of developmental regulators in order to establish repressive marks, H3K27me3 and H2AK119ub1, and suppress differentiation (Boyer et al., 2006). PRC2 components and Ring1B seem to be important during early stages of development, as depletion of Eed, Suz12, Ezh2 or Ring1B/A is letal due to gastrulation defects (reviewed in Landeira and Fisher, 2010). Apart from their roles in ES cell self-renewal and differentiation, PRCs are also important in the regulation of cell proliferation, senescence, and tumorigenesis (e.g. Jacobs et al., 1999, Varambally et al., 2002).

Initial studies described PRC1 as an inhibitor of RNAPII recruitment through a mechanism that prevented the access of SWI/SNF complexes to the TSS thereby preventing PIC assembly due to (compact) chromatin inaccessibility (Francis et al., 2004, Dellino et al., 2004). However, more recently, several studies have shown that for the great majority of genes, PRC-binding does not prevent the binding of the transcriptional machinery (Breiling et al., 2004, Stock et al., 2007, Chopra et al., 2009b), thus suggesting that PRCs act to inhibit transcription elongation, rather then RNAPII recruitment. The mechanism of interfering with transcription elongation is thought to involve H2A ubiquitnation (Zhou et al., 2008, Stock et al., 2007) and/or chromatin compaction (Eskeland et al., 2010). However, Stock et al. (Stock et al., 2007) showed that in ES cells, RNAPII-S5P complexes elongate through gene coding regions of PRC-repressed genes, and suggested that PRC1 does not inhibit transcription elongation, but instead may interact with the RNAPII CTD which possibly affects productive mRNA processing (Brookes and Pombo, 2009).

1.5.3 *Polycomb recruitment*

In *Drosophila*, Polycomb group proteins bind to cis-regulatory motifs termed polycomb group responsive elements (PREs) (Mulholland et al., 2003). However, in mammals, PREs or specific DNA sequences associated with the recruitment of PcG proteins have not been identified. Mechanisms of PRC recruitment to its targets in mammals is still under debate and is thought to involve more then one mechanism. For instance, the DNA binding protein and transcriptional repressor, YY1, is currently recognised as a mouse/human Polycomb protein (since it is a homologue of *Drosophila* PcG protein PhO and can replace it's function *in vivo*; Atchison et al., 2003), therfore providing a mechanism for PRC recruitment in vertebrates. Morover, noncoding RNAs have been shown to bind PRC2 and have been implicated in their recruitment.

In ES and CD4+ T cells, short ncRNAs (50 to 200nt) from PRC-targets were found to form stem-loop like structures that recruit Suz12 PRC2 subunit to their promoters (Kanhere et al., 2010). Upon activation of these PRC-bound genes through differentiation, short RNAs are depleted, thus demonstrating a mechanism by which transcription has repressive functions. The situation in *Drosophila* is different. It has been shown that transcription from a PRE leads to loss of Polycomb binding and

derepression of the associated genes in *Drosophila* (Bender and Fitzgerald, 2002, Hogga and Karch, 2002). In mammals several mechanisms of ncRNAs-mediated PRC recruitment have been found. One example is HOTAIR which, expressed from the *Hoxc* locus, recruits PRC2 to the *Hoxd* locus located in a different chromosome, thereby regulating PRC recruitment in *trans* (Rinn et al., 2007). Moreover, a genome-wide role for HOTAIR was shown in a human cancer cell line by overexpression of HOTAIR which led to global changes in PRC2 occupancy and H3K27me3, and in an increased cancer invasiveness and metastasis (Gupta et al., 2010). Other examples relate to PRC-targeting in *cis*, such as Xist involved in X chromosome inactivation (Zhao et al., 2008), and Kcnq1ot1 in the case of imprinted repression (Pandey et al., 2008). Many other long ncRNAs (lncRNAs) have been identified from human and mouse ES cells. In humans, ChIP-Chip experiments determined that about 20% of these lncRNAs are physically associated with PRC2 (Khalil et al., 2009). As such, it is possible that other mechanisms of ncRNA interactions to PcG proteins could exist.

1.5.4 *Chromatin bivalency*

In ES cells the great majority of PRC-targets display "bivalent" chromatin, characterised by the concomitant detection of H3K4me3 and H3K27me3 (Azuara et al., 2006, Bernstein et al., 2006; Figure 1.4). Bivalent genes are also enriched for H3K9ac (Azuara et al., 2006), and have CpG rich promoters (Mikkelsen et al., 2007), and recruit RNAPII to their promoters, which is able to elongate and produce transcripts that can be spliced, at least at the first intron (Stock et al., 2007). Bivalent genes were also shown to replicate early in S phase (Azuara et al., 2006). Interestingly, the histone variant H2AZ, typically associated with active genes, was found to co-localize with many PRC2 targets (Creyghton et al., 2008). H2AZ and H3K4me3 are mutually exclusive to DNA methylation (Cedar and Bergman, 2009, Zilberman et al., 2008), which is consistent with lack of DNA methylation at PRC-targets (Meissner et al., 2008). Genome-wide profiling of H3K4me3 and H3K27me3 identified thousands of bivalent genes in both mouse (Mikkelsen et al., 2007) and human (Pan et al., 2007, Zhao et al., 2007) ES cells, with ~50% of the targets being shared between mouse and human (Sharov and Ko, 2007). This suggests that bivalency is a conserved regulatory mechanism of gene control. Bivalency is thought to constitute a

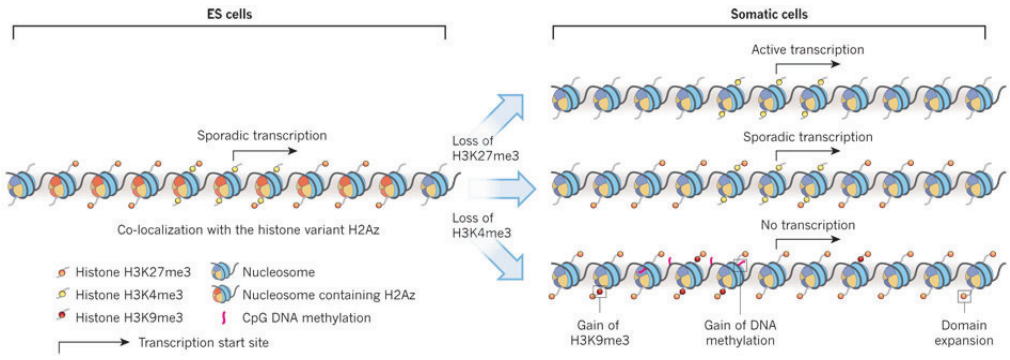


Figure 1.4: Schematic representation of chromatin at bivalent genes in ES cells and upon differentiation. Bivalent genes are marked with H3K4me3 and H3K27me3 and co-localize with the histone variant H2AZ. Upon differentiation H2AZ is removed and some bivalency is resolved. Loss of H3K4me3 can be accompained by other silencing pathways such as DNA methylation or H3K9me3. Image from Margueron and Reinberg [2011].

mechanism of "poising" genes for subsequent activation or repression upon differentiation, thus enabling ES cells to efficiently control gene expression during differentiation (Azuara et al., 2006, Bernstein et al., 2006; Figure 1.4).

Since these original genome-wide studies, several other studies questioned the specificity of bivalent domains to ES cells, as the co-association of H3K4me3 and H3K27me3 was also observed in differentiated cells, including NPCs (Mikkelsen et al., 2007), MEFs (Mikkelsen et al., 2007), lung fibroblasts (Pan et al., 2007), neurons (Mohn et al., 2008) and T cells (Barski et al., 2007, Roh et al., 2006), albeit at a lower frequency. Moreover, analysis of chromatin bivalency in *Xenopus* and *Drosophila* suggests a different mechanism of PRC repression. In *Xenopus* gastrula stage embryos, H3K4me3 and H3K27me3 do not co-occupy the same promoters (Akkers et al., 2009). In cultured *Drosophila* cells, the H3K4me3 HMT Ash1 was found at a low percentage (~5%) of PRC-target genes, and was suggested to reflect alternating states or spatial regulation, rather then true co-localisation (Schuettengruber et al., 2009).

Bivalent genes can resolve to a univalent state (either H3K4me3, active; or H3K27me3, repressed) or remain poised upon ES cell differentiation (Figure 1.4). It is not clear how bivalent marks are resolved, but it is thought that the mechanism involves histone HMTs and DNA methylation. For instance, PRC2 has been shown to recruit DNA DNMTs during ES cell differentiation which act to repress appropriate genes during lineage commitment (Vire et al., 2006). Bivalent chromatin might also be

transitory during differentiation. An example of such phenomena comes from *Nanog*, *Sox2* and *Oct4* pluripotency genes, which are active in ES cells but, before becoming repressed in differentiated cells, acquire a transient bivalent state (Pan et al., 2007).

1.6 RESEARCH AIMS AND OBJECTIVES

In this thesis I aimed to investigate the interplay between RNAPII and Polycomb repressor complexes in mouse ES cells. To this end, I integrated and analysed several genome-wide datasets obtained at different regulatory levels (epigenomics, transcriptomic, proteomics) and from different methodologies (ChIP-Seq, mRNA-Seq, Microarray). To elucidate RNAPII modifications genome-wide in mouse ES cells, and in particulary to characterise the PRC-repressed state, I examined RNAPII configuration at active, PRC-repressed and inactive genes. To further investigate the combinatorial association between RNAPII and Polycomb repressor complexes, I aimed to group genes according to their different combinations of PRC and RNAPII configurations using clustering techniques. Finally, I aimed to develop a software that can be used to analyse Next-Generation sequencing datasets and to assist in the interpretation of complex information, mainly through intuitive graphical representation of grouped data and statistical testing. The work presented in this thesis is mainly focused on the characterisation of the PRC-repressed state with respect to epigenetic and RNAPII key CTD modifications, and in the general elucidation of the "CTD code" in mouse ES cells.

1.7 THESIS OUTLINE

This thesis starts with a literature review (Chapter 1) of the state of the art on mechanisms of gene regulation, particularly focusing on current view of RNAPII CTD "code", transcription priming mechanisms, and on Polycomb-mediated gene repression in ES cells. Chapter 2 describes the methodology used, including a brief description of the experimental work and a more detailed description of bioinformatic methods employed in the work presented herein. Results are described in chapters 3 to 6. In Chapter 3, I start by analysing genome-wide ChIP-Seq mapping of RNAPII modifications, histone modification marks, and mRNA-Seq transcriptome

in ES cells, and review some of the current accepted paradigms of the CTD "code" at active and bivalent genes. In Chapter 4, I further integrate genome-wide maps of PRC subunits and PRC-instigated histone modifications in the analysis, to further investigate the interplay between RNAPII and PRCs. Here, I also describe the development of a new classification methodology used in this thesis to classify genes according to RNAPII and PRC-mediated histone modifications. In Chapter 5, I focus on the unexpected finding of subpopulations of "silent" and "active" PRC-targets, that are distinguished by the detection of RNAPII-S2P and full-length transcription at the "active" subpopulation. Here, I investigate the molecular basis of this antagonistic epigenetic profile, as well as formulate hypotheses of possible biological mechanisms leading to this finding. Finally, in Chapter 6, I describe a software that I have developed to aid in the analysis of ChIP-Seq data. I finish the thesis with a general discussion and future research prospects in Chapter 7.

Part II

METHODS

2

METHODS

2.1 EXPERIMENTAL METHODS

All wet laboratory experiments were performed by Emily Brookes or Kelly Morris. Their collaboration is indicated in each figure of the thesis.

2.1.1 *Cell culture*

Mouse ES-OS25 and ES-ERT2 cells were grown as previously described (Billon et al., 2002, Niwa et al., 2000, Stock et al., 2007). ES-ERT2 cells carry a tamoxifen-inducible, conditional knockout of the core PRC1 protein Ring1B, and are also homozygous null for the functional homologue Ring1A. For Ring1B conditional deletion, ES-ERT2 cells were treated for 24-72h with 800nM 4-hydroxy-tamoxifen (H7904, Sigma).

2.1.2 *Chromatin immunoprecipitation: single and sequential ChIP*

Fixed ChIP assays for RNAPII modifications, Ring1B and Ezh2 and native ChIP assays for histone modifications (H2Aub1, H3K36me3) were performed essentially as described previously (Stock et al., 2007, Szutorisz et al., 2005). Both assays were adapted for high-throughput analysis through the use of magnetic beads.

For Sequential ChIP (re-ChIP), fixed chromatin was prepared and the first immunoprecipitations set up using magnetic beads as for single ChIP. The second round of immunoprecipitations was then carried out, also following the standard protocol, adding magnetic beads and antibody directly. Enrichment was calculated relative to the original input (chromatin) using the same amount of DNA in the qPCR. To test for contamination of antibody remaining from the first

immunoprecipitation, we included a no-antibody control in the second round of immunoprecipitations as a negative control. As a positive control, we performed the first and second immunoprecipitations using the same antibody. The order of immunoprecipitations was also inverted to confirm specificity.

2.1.3 *ChIP Cross-Linking Vs Native ChIP*

Before the analysis of the ChIP chromatin, cross-linking of the DNA and proteins is often needed to fix the chromatin interactions between the antigen of interest and its binding site in the genome. We have used cross-linking ChIP protocol for the analysis of RNAPII modifications, Ring1B and Ezh2, while histone modifications (H3K36me3 and H2Aub1) were analysed using native ChIP. The reason why we have used different protocols is because histones normally do not need cross-linking as histones are generally strongly attached to the DNA. The cross-linking step has to be performed in very optimized conditions. Avoiding this step in a ChIP protocol can be advantageous since cross-linking can prevent the availability of epitopes or generate changes in the epitopes and thus reduce the relative abundance of the antigen. On the other hand, proteins with weaker affinities to the DNA, or that do not interact directly with the chromatin, need to be cross-linked in order for the ChIP to be efficient. Cross-linking is achieved through the use of formaldehyde, as the links formed can be reversed.

2.1.4 *ChIP validation by qPCR*

The isolated DNA from a ChIP experiment can be quantified using (real-time) quantitative PCR (qPCR). This technique depends on the use of primers which target specific DNA sequences. The material is then amplified and simultaneously quantified by measuring changes in fluorescence. This technique allows for the quantification of DNA (after ChIP enrichment) in a quicker and more cost effective way then whole genome sequencing. qPCR experiments were performed for the validation of samples sent for library preparation and of ChIP-Seq results.

2.1.5 *RNA purification and gene expression analysis by RT-PCR*

Total RNA was isolated using TRIzol (Invitrogen) extraction following the manufacturer's instructions and immediately treated with TURBO DNase I (Ambion). Treated RNA was reverse transcribed using random primers. The synthesized cDNA was diluted 1:10, and 2.5µl were used for reverse transcriptase PCR (RT-PCR).

2.1.6 *Illumina library preparation and sequencing: ChIP-Seq and mRNA-Seq*

ChIP-Seq libraries were prepared according to Illumina protocols with some modifications. Samples were PCR amplified prior to size selection. The mRNA-Seq library was prepared from total RNA isolated from ES-OS25 cells by TRIzol extraction, according to Illumina's instructions with some modifications. After polyA selection, ribosomal RNA was depleted using the Ribominus kit (Invitrogen), according to the manufacturer's instructions. Libraries were quantified by Qubit (Invitrogen) and qPCR, and library size was assessed by Bioanalyzer (Agilent). ChIP-Seq libraries were prepared at the WTCHG genomics facility, Oxford, in collaboration with Jiannis Ragoussis. mRNA-Seq libraries were prepared by Kelly Morris in collaboration with Jiannis Ragoussis.

2.1.7 *Antibody validation by ELISA*

ChIP is used for the selective enrichment of a chromatin fraction that contains a specific protein or protein modification of interest. Antibodies that recognize the antigen of interest are used to determine the amount of chromatin bound protein of modified chromatin. A list of all antibodies used in the experimental work presented in this thesis is summarised in Table 2.1.

The specificity of antibodies to RNAPII was characterised in detail. Enzyme Linked Immunosorbent Assay (ELISA) and western blotting are typically used to demonstrate that the correct antigen is detected by the antibody. In the case of RNAPII CTD modifications, ELISA was used to test whether modifications adjacent to the modification for which the antibody was originally raised can influence the antibody performance. ELISAs were performed for all RNAPII antibodies used with peptides

with different combinations of modifications in close proximity to the modified aminoacid of interest to test if antibody affinity could be inhibited or decreased. ELISAs were performed by Dirk Eick, Center of Integrated Protein Science, Munich; Naohito Nozaki, Institute of Technology, Yokohama, Tokyo; and Hiroshi Kimura, Graduate School of Frontier Biosciences, Osaka.

2.1.8 *Immuno-cryoFISH*

Co-localization of DNA loci with RNAPII 'transcription factories' was performed by Sheila Xie on ultrathin cryosections as described previously (Branco and Pombo, 2006, Ferrai et al., 2010).

2.2 HIGH-THROUGHPUT SEQUENCING AND ALIGNMENT

2.2.1 *ChIP-seq datasets*

A summary of ChIP-seq datasets is presented in Table 2.1. We generated ChIP-Seq data for the RNAPII modifications S5P, S7P, S2P and 8WG16, for the Polycomb protein Ring1B, and for the histone modifications H3K36me3 and H2Aub1. We used published ChIP-seq data for H3K4me3, H3K27me3 and core H3 (Mikkelsen et al., 2007), for Suz12 and Ezh2 (Ku et al., 2008); all of which were generated in mouse ES-V6.5 cells. For initial comparisons of our data with existing datasets, we used published Ring1B data (Ku et al., 2008), and RNAPII-S5P and -S2P data from Rahl et al. (Rahl et al., 2010), also generated in mouse ES-V6.5 cells. Libraries were sequenced using Illumina Sequencing Technology at the WTCHG genomics facility or the MRC-CSC Genomics laboratory using an Illumina Genome Analyser II, according to manufacturer's instructions.

2.2.2 *mRNA-Seq datasets*

We generated mRNA-Seq data after polyA+ RNA purification and ribosomal RNA depletion. We also analysed published mRNA-Seq data (SOLiD sequencing data; Cloonan et al., 2008). Gene expression datasets used in this study are summarized in Table 2.1. Libraries were sequenced using Illumina Sequencing Technology at the

MRC-CSC Genomics laboratory using an Illumina Genome Analyser II, according to manufacturer's instructions.

2.2.3 *Alignment of sequenced reads*

The following criteria were used to filter and align the DNA reads for further analysis: each tag was required (i) to pass the Illumina chastity filter; (ii) to match uniquely to the most recent mouse genome version (mm9) and (iii) to have at most two mismatches (thus defining a 'unique' DNA read). Allowing for mismatches is important since there could be sequencing errors or differences in the sequenced genome compared to the genome of reference (e.g. single-nucleotide polymorphisms, genome annotation errors) that prevent the DNA read to align perfectly to the reference genome. ChIP-Seq datasets were aligned using ELAND pipeline versions 1.3 or above and mRNA-Seq datasets were aligned with the ELAND RNA implementation of CASAVA 1.6.

Published ChIP-Seq datasets (fastq format) were re-mapped to the mm9 genome using Bowtie software, version 0.9.8.1 (Langmead et al., 2009). When aligning reads with Bowtie, one can choose to either ignore any read that maps to more then one location or to keep one of the hits among the multiple possibilities. To allow direct comparison with Illumina data (allowing 2 mismatches and unique reads only), the option "-m 1" was used. Specifying "-m 1" instructs Bowtie to discard any read with more than 1 reportable alignment.

Table 2.1: List of ChIP-Seq and mRNA-Seq datasets used in this thesis.

Seq dataset	Antibody	Dataset source	Cell type	Mapped reads (millions)
S5P	CTD4H8 (MMS-128P, Covance)	–	ES-OS25	22
S5P	ab5131 (Abcam)	Rahl et al. [2010] (GSM515662)	ES-V6.5	4
8WG16	8WG16 (MMS-126R, Covance)	–	ES-OS25	24
S7P	4E12 (Chapman et al., 2007)	–	ES-OS25	11
S2P	H5 (MMS-129R, Covance)	–	ES-OS25	33
S2P	H5 (ab24758, Abcam)	Rahl et al. [2010] (GSM515663)	ES-V6.5	11
Ring1B	A-20 (Atsuta et al., 2001)	–	ES-OS25	28
Ring1B	A-20 (Atsuta et al., 2001)	Ku et al. [2008] (GSM327669)	ES-V6.5	4
Ezh2	39103, Active Motif	Ku et al. [2008] (GSM327668)	ES-V6.5	9
Suz12	ab12073 (Abcam)	Ku et al. [2008] (GSM327667)	ES-V6.5	13
H2Aub1	E6C5 (Upstate)	–	ES-OS25	18
H3K27me3	07-449 (Upstate)	Mikkelsen et al. [2007] (GSM307619)	ES-V6.5	8
H3K4me3	ab8580 (Abcam)	Mikkelsen et al. [2007] (GSM307618)	ES-V6.5	9
H3K36me3	13C9 (Rechtsteiner et al., 2010)	–	ES-OS25	23
H3	ab1791 (Abcam)	Mikkelsen et al. [2007] (GSM307624)	ES-V6.5	7
mRNA-Seq (this study)	–	–	ES-OS25	93
mRNA-Seq	–	Cloonan et al. [2008] (SRA000306)	ES-W9.5	8

2.3 PROCESSING OF GENOME-WIDE SEQUENCING DATA

Subsequently to the alignment of the DNA reads to the genome, the processing of the ChIP-Seq and mRNA-Seq data is followed by the analysis of read density patterns. The first step is to convert the alignment files from the typical tags-to-positions to count-per-position data (i.e. bedGraph files). Subsequently, sequencing of depth at specific windows of interest can be calculated and used to draw aggregate profiles that summarize the pattern of enrichment at given genomic window for a group of genes.

2.3.1 *Generation of bedGraph Files*

The bedGraph file format contains genome-wide nucleotide-resolution information of read coverage. 'bedGraph' files are a file type that can be directly uploaded into the University of California Santa Cruz (UCSC) Genome Browser (Kent et al., 2002) which allow visual inspection the sequencing depth and coverage at any genomic location of interest.

2.3.2 *Computation of sequencing depth of coverage at genomic intervals of interest*

Data for each ChIP-Seq library are presented in windows covering gene regions of interest. Read "depth of coverage" was computed by calculating the area under the coverage curve (i.e. number of sequenced nucleotides that overlap the window of interest). Box plots displaying the median, upper and lower quartiles and the range of the data were generated by calculating the depth of coverage in 2kb windows of interest located either relative to the TSS or TTS of genes. R software (R Development Core Team, 2010) was used to generate heatmaps, average profile plots and box plots. Python scripts were used to generate "depth of coverage" counts.

2.3.3 *Data visualisation in the log scale*

Log_{10} transformation was applied before plotting heatmaps, box plots and correlation plots. As values equal to zero are meaningful in ChIP-Seq and mRNA-seq data, a constant (+1 and +0.0001, respectively) was added to all data points before log

transformation or before length normalisation, for ChIP-Seq depth of coverage or mRNA-Seq FPKM levels, respectively, unless otherwise stated. ChIP-Seq levels were quantified using "depth of coverage" at each genomic window of interest (relative to the TSS or TTS of genes). Therefore, within all windows, values typically range from 0 to five orders of magnitude and adding the constant 1 did not affect the overall distribution of the data values. FPKM levels ranged from 0 to three orders of magnitude and thus the constant added had to be chosen accordingly.

2.3.4 *Visualisation of the distribution pattern of ChIP-Seq reads in gene 5′ and 3′ regions*

In order to access the abundance of ChIP-Seq reads at gene regions of interest, depth of coverage was calculated in non-overlapping 10bp windows (to produce average ChIP-seq profiles) or in 100bp windows (to produce heatmaps).

To analyse the distribution pattern of read density near the 5′ end or 3′ end, genes were aligned according to their TSS or TTS. To produce average ChIP-Seq profiles, the mean value at each window position was calculated. Line graphs or heatmaps were drawn to visualize the profile of read density in R. A region of up to 10kb around the TSS or TTS (5 kb upstream and 5 kb downstream), was used in the analysis of each gene.

2.3.5 *FPKM calculation*

FPKM levels were calculated for our own mRNA-Seq library and for mRNA-Seq datasets published in Cloonan et al. [2008] using Tophat version 1.0.13 (Trapnell et al., 2009) and Cufflinks version 0.8.2 (Trapnell et al., 2010), to allow detection of reads crossing exon-exon junctions and calculation of FPKM levels per transcript.

2.4 ANALYSIS OF PUBLISHED ANNOTATION DATASETS

2.4.1 *Microarray datasets and gene expression analyses*

Details of all genomic annotations used in this study are listed in Table 2.3.

For the comparison of mRNA-Seq expression data with microarray datasets, we used published gene lists of expression values directly downloaded from two studies in which the authors have used GeneChip microarrays (Mikkelsen et al., 2007, Sene et al., 2007).

For the analysis of Ring1B knockout data, Affymetrix microarray mRNA profiles for mouse ES-ERT2 cells prior to and 48h after Ring1B deletion (Ring1A/B-dKO) were obtained from published CEL files (GSM265042, GSM265043; Endoh et al., 2008). When multiple probes matched to a unique gene position the average of all $\log_2$ fold-change intensities was taken as the expression level for that gene.

For the analysis of differentiation data, raw Affymetrix microarray expression datasets from mouse ES cells in undifferentiated and differentiated (2, 4, 6, and 8 days after LIF withdrawal) states were obtained from a published CEL files (GSE19165; Shen et al., 2009). Raw Affymetrix data was normalized and annotated using the GCRMA and mouse4302.db Bioconductor packages, respectively (Gentleman et al., 2004, Wu et al., 2004). To visualize differentiation microarray data we used MultiExperiment Viewer (Saeed et al., 2003). The bioinformatic analysis of differentiation data from Shen et al. [2009] was performed by Tom Carroll, MRC-CSC, London.

2.4.2 *Other genome-wide annotation datasets*

The analyses presented in this study rely on the use of several other published annotations which are summarised in Table 2.3. Namely, we have used proteome (Graumann et al., 2008, Van Hoof et al., 2006), DNA methylation (Fouse et al., 2008) and transcription factor published datasets (Chen et al., 2008, Kim et al., 2005). In all cases, the data was downloaded as gene lists and used directly after converting between gene symbols. In this thesis, I have used UCSC IDs as the primary gene identification ID. Gene IDs were converted from RefSeq or Entrez to the UCSC IDs using mm9 conversion tables (Karolchik et al., 2003), namely *mm9.kgXref.refseq* and *mm9.refLink.locusLinkId*, respectively, and R software.

Table 2.3: Sources of genomic annotations downloaded from selected published papers

Dataset and Source	Cell type	Additional Information
mRNA Microarrays; Mikkelsen et al. [2007]	ES-V6.5	cDNA was generated from poly(A)+ mRNA and hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 Array.
mRNA Microarrays; Sene et al. [2007]	ES-V6.5	cDNA was generated from poly(A)+ mRNA and hybridized to Affymetrix GeneChip Mouse Genome 430A and 430B set.
mRNA Microarrays; Ring1A/B knockout; Endoh et al. [2008]	ES-ERT2	Ring1A/B double-knockout (dKO) was generated by OHT treatment during 48 hours. Genome-wide mRNA analysis were performed using Affymetrix GeneChip Mouse Genome 430 2.0 arrays.
mRNA Microarrays; ES cell differentiation; Shen et al. [2009]	ES-J1	Wild-type ES cells were differentiated after LIF withdrawal and mRNA was extracted from day 0, 2, 4, 6 and 8. Genome-Wide mRNA analysis was performed using Affymetrix GeneChip Mouse Genome 430 2.0 arrays.
Total ES proteome; Van Hoof et al. [2006]	ES-D3	Proteins were extracted from ES cells and identified by mass spectrometry.
Total ES proteome; Graumann et al., 2008	ES-R1	Proteins were extracted from ES cells and quantified by stable isotope labeling by amino acids in cell culture (SILAC).
DNA methylation (mDIP-Chip); Fouse et al. [2008]	ES-J1	Genome-wide mapping of DNA methylation patterns were obtained by coupling methylated DNA immunoprecipitation with DNA microarray chip technology (mDIP-Chip). DNA was immunoprecipitated with an antibody to 5-Me-cytosine. Microarray hybridisations were performed on the Agilent Whole Genome microarrays (G4122A).
TF ChIP-Seq; Chen et al. [2008]	ES-E14	Binding sites for thirteen transcription factors were mapped using commercially available antibodies and ChIP-Seq. To associate binding sites with genes, an association score for each pair of gene and TF was computed based on the relative distance to the TSS of the gene.
TF ChIP-ChIP; Kim et al. [2008]	ES-J1	*In vivo* binding sites for nine transcription factors were obtained by biotinylation mediated ChIP (bioChIP). Microarray hybridisations were performed on the Affymetrix GeneChip Mouse promoter 1.0R arrays.

2.5 CHOICE OF A NON-REDUNDANT GENOME

UCSC Known Genes annotation track contains gene predictions based on data from RefSeq, Genbank, and UniProt (Karolchik et al., 2003). The mouse genome dataset (mm9 version, release date March 2007) contains 49409 annotated transcripts, many of those corresponding to splice variants of the same gene.

To build a non-redundant gene set, I used the information from *mm9.KnownIsoforms.clusterid* and *mm9.KgXref tables* to link annotated transcripts into clusters that represent genes. The initial dataset comprised a total of 18860 gene clusters, where each cluster was represented by one or more RefSeq genes. Genes belonging to random chromosomes were discarded. I selected the 'representative' UCSC gene from each cluster according to the following criteria (schematized in Figure 2.1):

1. For each gene cluster, we selected as representative the isoform with the highest S5P enrichment in a 2kb window centered at the isoform TSS;
2. For the gene clusters that were still represented by more then one isoform, we selected the isoform with the highest S2P enrichment in a 4kb window centered at the isoform TTS;
3. For clusters that survived the previous criteria, the representative gene was selected by considering the canonical transcript as annotated in the *mm9.knownCanonical.transcript* UCSC table. According to this process, within a cluster, the gene with the highest number of coding bases is chosen as the representative canonical gene;
4. Finally, for complete duplicate entries (based on the previous criteria), we selected a random isoform as the representative for that gene cluster.

This process produced a list of 18860 unique UCSC Known Gene IDs, and their corresponding RefSeq IDs.

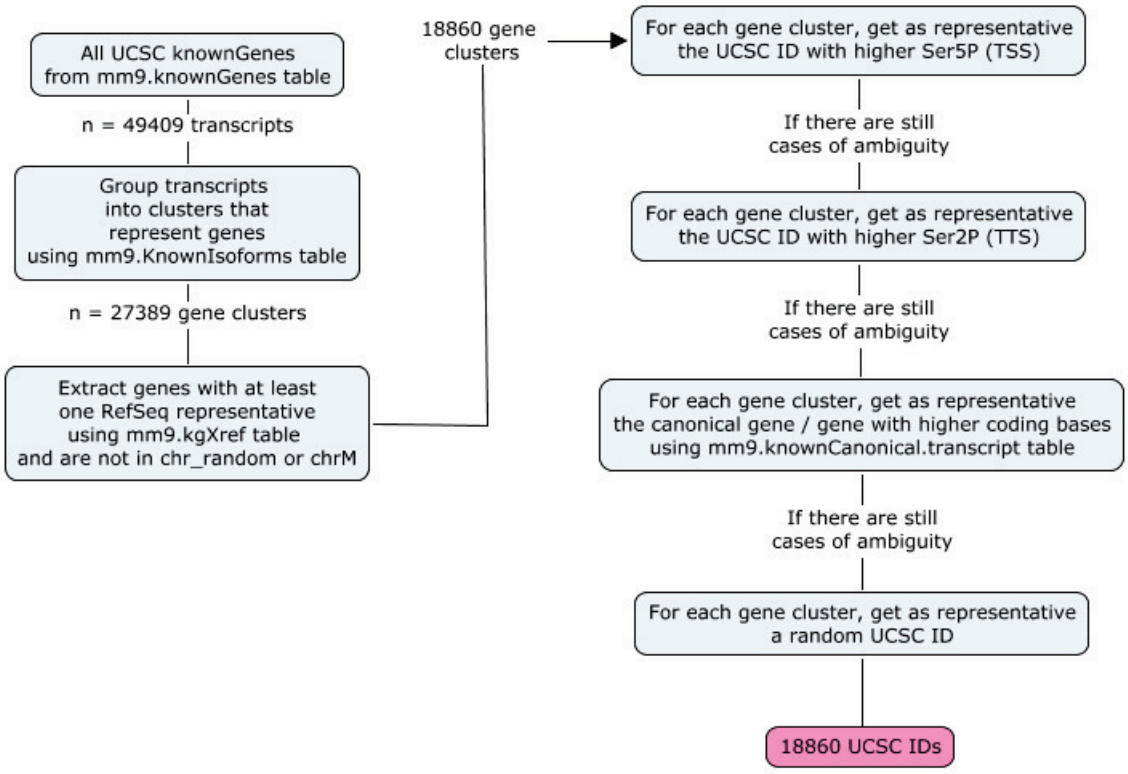


Figure 2.1: Scheme of the methodology used to build a candidate gene set for downstream analyses.

2.6 GENE ONTOLOGY ANALYSIS

Analysis of GO functional enrichment was performed using the Fisher's exact test implemented in the topGO Bioconductor package (Alexa et al., 2006) and the GO term to Entrez gene ID annotation provided within the org.Mm.eg.db Bioconductor annotation database (Version 2.4.1; Gentleman et al., 2004). Enrichment of GO terms for each gene cluster was tested using all genes within every cluster as a reference set. Significance was identified for GO terms with p-values less than 0.05. In order to capture GO terms of higher order annotation, the resulting enriched GO terms were filtered to their significant ancestors which were highest within the GO graph. The method of capturing the GO term of higher order was developed by Tom Carroll.

2.7 KEGG PATHWAY ANALYSIS

Annotation for KEGG pathways and their associated genes were retrieved from the KEGG FTP site (ftp://ftp.genome.jp/pub/kegg/; Kanehisa and Goto, 2000). Enrichment of KEGG pathways by group members was assessed by hypergeometric testing using the R Stats package and false discovery rates calculated using the R Multtest package (Venables and Smith, 2003).

2.8 CLASSIFICATION SCHEME

To investigate the combinatorial association of each gene with RNAPII or PRC-mediated histone modifications, we classified each gene as positive or negative for each modification at regions of interest. We integrated the "depth of coverage" within the selected windows of interest for each modification across all RefSeq genes to capture local read density. Each gene was characterised by the sequencing depth at 2kb windows defined relative to either the TSS or TTS, depending on the factor analysed. To convert depth of coverage (i.e. continuous values) into binary classifications, we used an expectation maximization algorithm (*optim* function in R) to fit a two-component Gaussian mixture distribution to the distribution of sequencing depth data values. A Gaussian curve on the left was used to model the 'noise' distribution and a lognormal curve on the right models the distribution of the 'real signal'. Ultimately, the curve fitting approach allowed the identification of a threshold at a given false discovery rate (FDR). The FDR is the quotient between the expected proportion of false positive detections and the total number of detections. We used an FDR of 1%, thus setting the threshold to a value where the proportion of 'noise' corresponds to 1% of the total area of the mixture on the right of the threshold, ensuring that the number of false positive detections will be small. The threshold was then used to categorise the continuous values into binary 'positive' or 'negative' (1,0) events that denote, respectively, the presence or absence of a given histone mark or RNAPII (either at TSS or TTS region). Gene classification methodology was developed in collaboration with Sarah Teichman and Daniel Hebenstreit, MRC-LMB Cambridge, and recently published as a webtool (Hebenstreit et al., 2011).

2.9 HIERARCHICAL CLUSTERING

The input matrix was composed of 15404 RefSeq genes and seven binary variables. The binary variables corresponded to 0/1 values which denoted the presence/absence of a given modification at the analysed gene region of interest. The gene regions of interest were chosen prior to the binary classification (either 2kb window centered at the TSS, 2kb window upstream the TTS or a 2kb window downstream the TTS). The seven variables (and respective gene windows) used in the cluster were: S5P (TSS), 8WG16 (TSS), S7P (TSS), S5P (upstream TTS), S2P (downstream TTS), H3K27me3 (TSS) and H2Aub1 (TSS). All pairwise dissimilarities in the data matrix were computed in R statistical software using the Gower coefficient (Gower, 1971); hierarchical clustering was calculated using the average linkage method and the function *hclust* in R.

2.10 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis was performed in R using the function *prcomp* and the same data matrix used for the hierarchical clustering (15404 x 7), but using the continuous values instead of the binary classification. Continuous values corresponded to sequencing depth at each gene window considered in the analysis (either 2kb TSS windows or 2kb TTS windows). Since the number of sequenced reads is different between ChIP samples, the data matrix had to be normalised prior to the PCA analysis. The method used is known as "standardization" and involves centering and scaling each column by subtracting the mean and dividing by the standard deviation. This was done by using *scale* function in R.

2.11 CPG AND GC CONTENT ANALYSIS

2.11.1 *Calculation of percentage CpG and GC*

DNA sequences covering -1 to +1kb of the TSS of each gene were downloaded from UCSC table browser (Kuhn et al., 2009). Percentage CpG and percentage GC were calculated for each 2kb sequence using the formula:

$$\%CpG = \frac{(number\ CpG)\ x2}{sequence\ length} \qquad \%GC = \frac{number\ of\ C + number\ of\ G}{sequence\ length}$$

2.11.2 *CpG O/E calculation and distribution pattern*

In order to visualise the distribution profile of CpG relative abundance at gene 5′ end regions, I calculated the CpG "observed over expected" ratio (O/E) in windows of length 200bp with an overlap of 150bp between windows. CpG O/E was calculated using the formula:

$$CpG\ (O/E) = \frac{number\ of\ CpG}{number\ of\ C\ x\ number\ of\ G}$$

To produce average CpG O/E profiles, genes were aligned according to their TSS and the mean value at each window position was calculated. A region covering -3 to +3kb of the TSS of each gene was used in the analysis. The distribution pattern of %GC across the same regions was calculated as described, but instead of CpG O/E, the %GC was used.

2.12 GSEA ANALYSIS

Gene-set enrichment analysis (GSEA) were performed using GSEA software (Subramanian et al., 2005). Given a ranked gene list and a list of genes of particular interest, GSEA tests whether the genes of interest are randomly distributed along the ranked list or particularly found near the edges, with the latter indicating an association of high/low scores. GSEA was originally built to interpret gene expression data, however the statistical method can be used to test if any *a priori* set of genes are significantly enriched for a particular biological state.

2.13 SEQGI SOURCE CODE IMPLEMENTATION

SeqGI (Chapter 6) is a software developed to aid analyses and comparison of ChIP-Seq data. Its source code consists of three components: (i) OGRe script, a standalone Python script used to calculate read density at user defined genomic windows; (ii) the graphical user interface (GUI) written in Python; (iii) and a set of statistical tests written in R to analyse and compare count-type data.

SeqGI home page is available online at http://seqgi.sf.net. All source code is available to download at http://sourceforge.net/projects/seqgi/.

2.13.1 *Computation of overlapping genomic intervals via OGRe.py*

I developed OGRe.py based on the original concept of "Fjoin" algorithm published by Richardson [2006]. OGRe script was written in Python (http://www.python.org/). Python interpreter is called within SeqGI to run OGRe, which can also be used as a standalone script. When used within SeqGI, OGRe output is integrated with SeqGI for visualisation and statistical testing.

2.13.2 *SeqGI graphical user interface*

The GUI was written in Python using wxPython toolkit (http://www.wxpython.org/) and the GUI designer wxGlade (http://wxglade.sourceforge.net/). Visualisation tools were implemented using matplotlib python 2D plotting library (http://matplotlib.sf.net/).

2.13.3 *Statistical analyses within SeqGI*

Statistical analyses were implemented in R and integrated within the workflow of SeqGI. R version 2.9.12 is distributed with SeqGI. SeqGI communicates user instructions to R using "subprocess" python standard library, which performs the test and outputs the result back to SeqGI. Implementation of R scripts and methodology for the statistical analysis is being developed in collaboration with Tom Carroll.

Part III

RESULTS

3

GENOME-WIDE MAPPING OF RNAPII AND TRANSCRIPTOME ANALYSIS IN MURINE ES CELLS

3.1 RESEARCH MOTIVATION

NGS technologies are quickly being used to explore biological questions in a genome-wide fashion. Chromatin-immunoprecipitation followed by sequencing (ChIP-Seq) is among the NGS applications widely used to investigate the profile of DNA-binding proteins or other chromatin-related features, such as histone modifications.

My aim in this project was to provide a genome-wide characterization of RNAPII CTD modifications within the mouse ES cell genome using bioinformatic analysis of ChIP-Seq data. In particular, I aimed to examine the CTD modifications at genes with different epigenetic states, from the most expressed genes to bivalent and the least expressed genes in the genome. Moreover, to gain insights into the function of "poised" RNAPII complexes, I aimed to characterise the transcriptional status of bivalent genes using for this propose publicly available expression datasets and our own deep mRNA-Seq profiling of the ES transcriptome.

The experimental work presented in this chapter was carried out by E. Brookes. All ChIP-Seq libraries were prepared at the WTCHG genomics facility, Oxford, in collaboration with Jiannis Ragoussis. mRNA library was generated by K.J.Morris.

3.2 STATE OF THE ART

3.2.1 *RNAPII CTD "code"*

In mammals, the RNAPII CTD consists of 52 heptapeptide repeats of the consensus sequence Y1-S2-P3-T4-S5-P6-S7 (Corden et al., 1985). When RNAPII is attached to the chromatin, a yet unknown number of CTD repeats become phosphorylated, at least, on three serine residues, S5, S2 and S7. CTD post-translational modifications take place at different stages of the transcription cycle. Several studies have shown that RNAPII with particular CTD phosphorylation patterns (CTD "code") act as a recruiting platform for different HMTs and RNA processing machinery, integrating chromatin modification with RNA processing and transcription (Hirose and Ohkuma, 2007). Serine 5 phosphorylation (S5P) is associated with transcription initiation, capping (Ho and Shuman, 1999, Rodriguez et al., 2000) and recruitment of H3K4 methylases (Milne et al., 2005, Ng et al., 2003), while S2P correlates with elongation, splicing, polyadenylation and recruitment of H3K36 methylases (Krogan et al., 2003, Xiao et al., 2003, Li et al., 2003). S7P has been more recentely addressed, and was found to be present at promoters and gene bodies of actively transcribed genes in yeast (Mayer et al., 2010, Tietjen et al., 2010, Kim et al., 2010a).

3.2.2 *RNAPII at bivalent genes in ES cells*

RNAPII is recruited to many inactive genes in ES cells (Guenther et al., 2007, Kim et al., 2005) and in the early *Drosophila* embryo (Zeitlinger et al., 2007), including polycomb-repressed genes (Chopra et al., 2009b, Stock et al., 2007). In mouse and human ES cells, Polycomb-repressed genes contain chromatin modifications that are associated with gene activity (H3K4me3) and repression (H3K27me3, H2Aub1) modifications. The occurrence of H3K4me3 and RNAPII at the promoter regions of inactive genes was interpreted to prime genes for subsequent activation (Azuara et al., 2006, Bernstein et al., 2006). Bivalent genes are enriched in developmental specific functions (Bernstein et al., 2006), are replicated early (Azuara et al., 2006), have CpG-rich promoters (Mikkelsen et al., 2007), and are transcribed at low levels in ES cells (Stock et al., 2007).

Our laboratory has demonstrated the presence of RNAPII "poised" at a panel of developmental regulator, bivalent genes in mouse ES cells, despite their PRC-dependent repression (Stock et al., 2007). The population of poised RNAPII is characterized by an unusual pattern of CTD phosphorylation, with high levels of phosphorylation at S5, no detectable S2 phosphorylation, and an unexpected lack of recognition by an antibody which recognizes the unphosphorylated S2 residues (8WG16). The importance of a CTD modification "code" in dictating epigenetic and expression states, suggests that the unusual set of CTD phosphopatterns of RNAPII complexes at PRC-repressed genes may be important in the silencing of these genes, while keeping them poised for activation (Brookes and Pombo, 2009). More extensive analysis became necessary to further explore combinations of CTD modifications on a global scale in ES cells. We used ChIP-Seq to generate genome-wide maps of RNAPII modifications other publicly available datasets to further expand our knowledge of detailed RNAPII configurations at active, PRC-repressed and inactive genes in the mouse ES cell genome.

3.3 RESULTS

3.3.1 *Generation of genome wide transcriptome data*

To obtain genome-wide coverage of the mouse transcriptome, cDNA was generated from mouse OS25 ES cells and subsequently sequenced using Illumina sequencing in four consecutive lanes. mRNA-Seq library was prepared by K. Morris. Reads were aligned against the UCSC mm9 genome (Karolchik et al., 2003), with the ELAND RNA implementation of CASAVA 1.6. A total of 107 million high quality sequenced reads (i.e. reads that passed Illumina's chastity filter) were generated, of which 87% (93 million) mapped to the mouse genome Table 3.1.

Table 3.1: Number of mapped reads from mRNA-Seq dataset. The percentage of mapped reads correspond to the number of reads that map to the mm9 genome out of the total number of reads that pass Illumina's quality chastity filter.

Run	Mapped reads (millions)	% of mapped reads
lane 1	20.4	75.6%
lane 2	24.3	89.0%
lane 3	24.2	89.5%
lane 4	24.0	88.8%

Comparison of the four technical replicates revealed that the method is highly reproducible, with Spearman's rank correlation coefficient ranging from 0.76 to 0.77 when comparing the location of aligned reads between pairs of libraries (Figure 3.1).

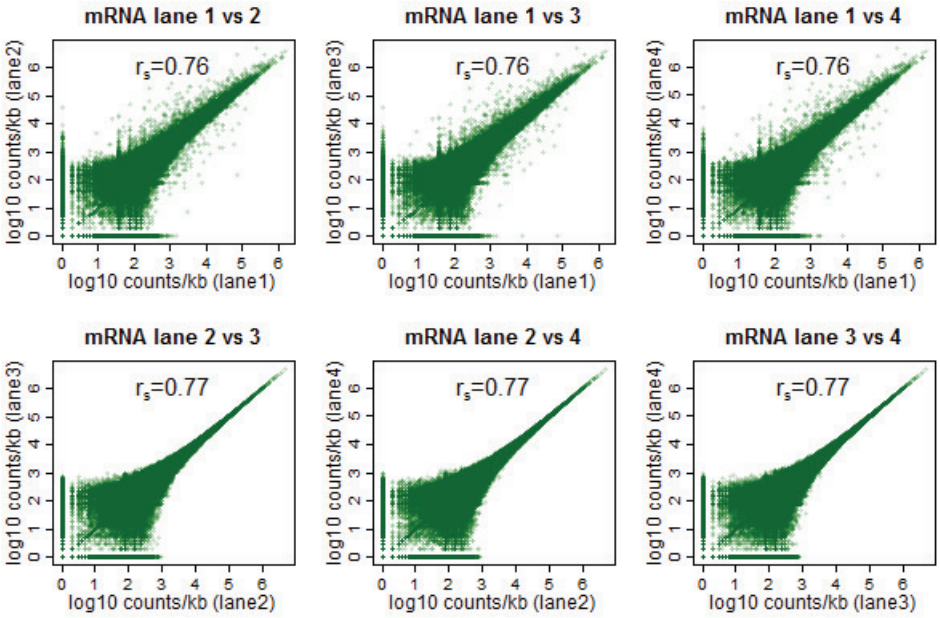


Figure 3.1: Comparison of read counts between independent sequencing lanes of the mRNA library. Counts were calculated in 1kb windows covering the entire mouse genome and log10 is plotted for each pair of technical replicates. To avoid taking the log of 0, the constant 1 was added to raw data. There is a strong positive correlation between sequencing lanes. Spearman's rank correlation coefficient is depicted in each scatter plot (r_s).

To validate the uncommitted state of the OS25 ES cells used to prepare the libraries, I inspected expression levels across pluripotency and differentiation specific marker genes (Figure 3.2). A high number of sequenced tags overlapping pluripotency genes *(Oct4, Sox2, Nanog)* was identified, appropriately enriched at the exonic regions, and a low number of tags at differentiation marker genes *(Sox17, Gata4, Gata6)*. Thus, the cultured OS25 cells represent a population of undifferentiated ES cells.

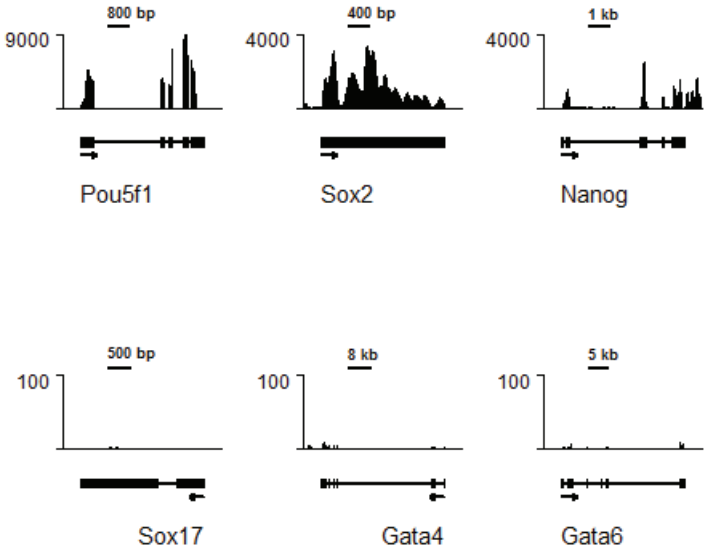


Figure 3.2: Inspection of mRNA-Seq profiles and levels across single genes. *Oct4*, *Sox2* and *Nanog* are expressed in pluripotent cells, while *Sox17*, *Gata4* and *Gata6* represent markers for differentiation. Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

3.3.2 *Generation of genome-wide profiles of RNAPII phosphorylation*

To begin exploring the CTD code, the occupancy profiles of three major CTD marks was examined across the mouse genome.

CTD phosphorylation was mapped by ChIP-Seq using a panel of four well characterised antibodies (Xie et al., 2006, Stock et al., 2007, Ferrai et al., 2010): 4H8 antibody recognizes S5P residues; 4E12 recognizes S7P residues; H5 binds to S2P residues, and finally 8WG16 which targets unphosphorylated S2 residues. The specificity of all RNAPII antibodies was confirmed by ELISAs (Brookes, de Santiago et al., 2011; submitted). As a control DNA immunoprecipitated in the absence of antibody (mock IP) was sequenced. Immunoprecipitated DNA was prepared by E. Brookes using wild-type mouse OS25 ES cells and libraries were sequenced in two to three runs for each biological sample (RNAPII, mock IP and H3K36me3 samples) using Illumina's Genome Analyzer II at the WTCHG genomics facility, Oxford.

To integrate CTD phosphorylation with chromatin histone modification patterns, we investigated H3K4me3 and H3K36me3 genome-wide profiles. H3K36me3 genome-wide maps were obtained in our laboratory from Illumina sequencing of a Native ChIP sample, while H3K4me3 datasets were obtained from Mikkelsen et al. [2007] in mouse V6.5 ES cells, and downloaded from GEO database (GSM307618). Sequenced reads were subsequently aligned to the mm9 assembly of the mouse genome using the Illumina alignment software Eland for our ChIP-Seq samples or Bowtie alignment software (Langmead et al., 2009) for publicly available datasets, as described in Chapter 2 (i.e. allowing at most two mismatches in each alignment, and ignoring reads that align to multiple genomic locations). On average, between all sequencing runs, 73% of reads mapped to a unique genomic location. Possible reasons for reads not mapping to the genome include the presence of higher sequencing errors, insertions/deletions, polymorphisms, and reads mapping to multiple repetitive regions (Leleu et al., 2010). A summary of the number of DNA reads sequenced at each sequencing run, and their mapping to the mouse genome is presented in Table 3.2.

Table 3.2: List of ChIP-Seq datasets used in this study. The percentage of mapped reads correspond to the number of reads that map to the mm9 genome out of the total number of reads that pass Illumina's quality chastity filter.

ChIP-Seq dataset	Antibody	Nr of runs	Mapped reads (millions)	% of mapped reads (Illumina alignment)
S5P	CTD4H8 (MMS-128P, Covance)	2	lane 1: 10.9 lane 2: 11.1	lane 1: 80.9% lane 2: 81.7%
8WG16	8WG16 (MMS-126R, Covance)	2	lane 1: 7.0 lane 2: 17.4	lane 1: 77.7 % lane 2: 74.3%
S7P	4E12 (Chapman et al., 2007)	2	lane 1: 5.4 lane 2: 5.8	lane 1: 58.7% lane 2: 58.3%
S2P	H5 (MMS-129R, Covance)	3	lane 1: 5.4 lane 2: 7.2 lane 3: 20.5	lane 1: 75.7% lane 2: 77.9% lane 3: 77.5%
Control	No antibody	2	lane 1: 5.9 lane 2: 5.7	lane 1: 58.6% lane 2: 57.8%
H3K36me3	13C9 (Rechtsteiner et al., 2010)	2	lane 1: 10.4 lane 2: 12.8	lane 1: 84.1% lane 2: 84.9%
H3K4me3 (Mikkelsen et al., 2007)	ab8580 (Abcam) (GSM307618)	6	total: 9.2	

Before combining sequencing reads from different lanes, I calculated the sequencing depth across sliding windows of 1kb length covering the whole genome and correlated these levels between each pair of lanes. For each library, counts are highly correlated across lanes (Spearman's correlation ranges between 0.59 and 0.78; Figure 3.3) confirming the technical reproducibility and accuracy of Illumina sequencing data. Errors between ChIP-Seq lanes are greater then between mRNA sequencing lanes especially between windows with lower counts that do not correlate between technical replicates of ChIP libraries. These less well correlated windows are probably reflecting higher library complexity in Chip-Seq datasets that sample the whole genome and different number of mapped reads.

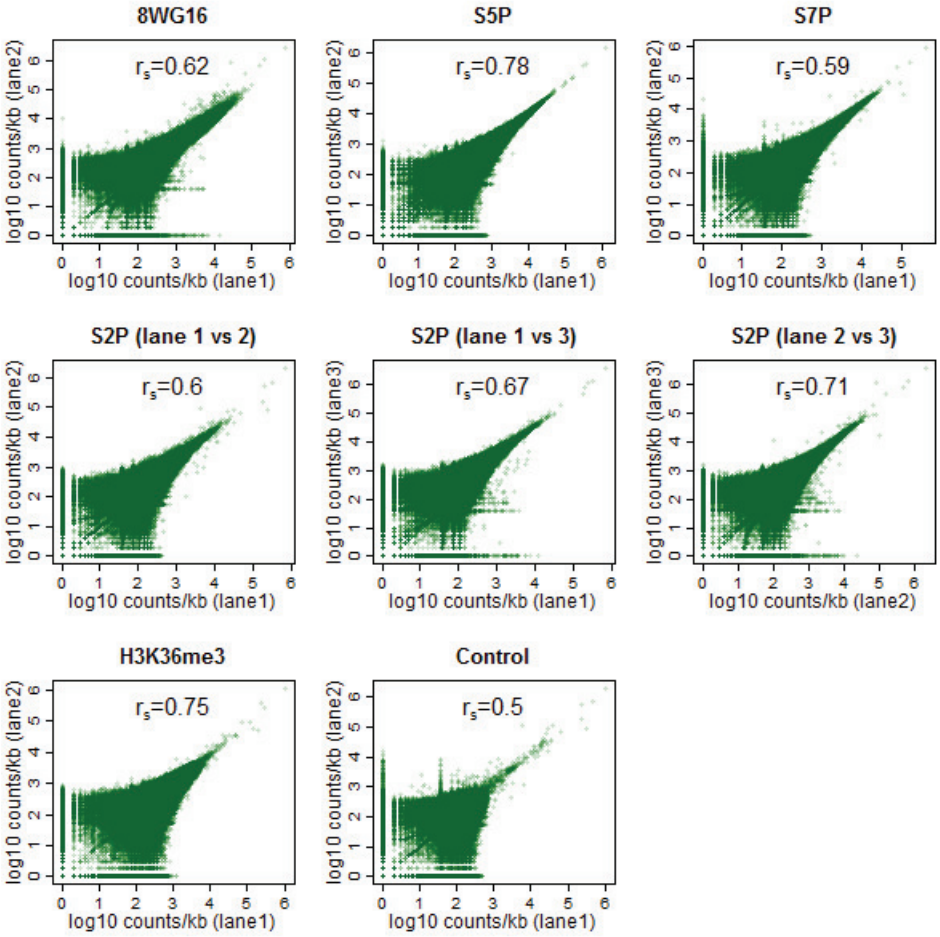


Figure 3.3: Reproducibility of sequencing replicates of the same ChIP-Seq library. In each panel, the log10 read counts for each 1kb window are plotted for each pair of technical replicates: RNAPII-S5P, -8WG16, -S7P, -S2P, H3K36me3 and mock IP (control) libraries. To avoid taking the log of 0, the constant 1 was added prior to taking logs. Spearman's rank correlation coefficient is depicted in each scatter plot (r_s). There is a strong positive correlation between independent sequencing lanes. Counts correspond to the number of read-nucleotides per 1kb window.

3.3.3 *Validation of the ChIP-Seq data by quantitative real-time PCR*

To validate the quality of ChIP-Seq experiments, the profile of ChIP-Seq enrichment was compared with qPCR analysis at single genes. ChIP-Seq and mRNA-Seq profiles (Figure 3.4) and qPCR profiles (Figure 3.5) were obtained for a panel of active, PRC-repressed and silent genes for which the binding of RNAPII phosphoforms (S5P, 8WG16 and S2P) has been well documented (Stock et al., 2007). At active genes (*ß-actin*, *Oct4*, *Sox2*), we detect high expression levels with appropriate enrichment of mRNA-Seq tags over spliced exons; S5P, 8WG16 and S7P peak at the promoter, while

S2P increases throughout the genes' coding region and peaks downstream of the TTS. Silent genes (*Myf5*) or genomic regions (5'HS2), show no or little enrichment for RNAPII modifications. Finally, single gene analyses by qPCR (Figure 3.5) show that S5P levels at PRC-repressed genes (*Atoh1*, *Msx1*, *Hoxa7*, *Gata4*) can be as high as at an active gene; however S2P is not detectable above background levels, confirming the presence of a novel pattern of CTD phosphorylation at PRC-repressed genes (Stock et al., 2007). ChIP-Seq profiles are comparable to qPCR results demonstrating, for the panel of selected genes, known RNAPII binding patterns.

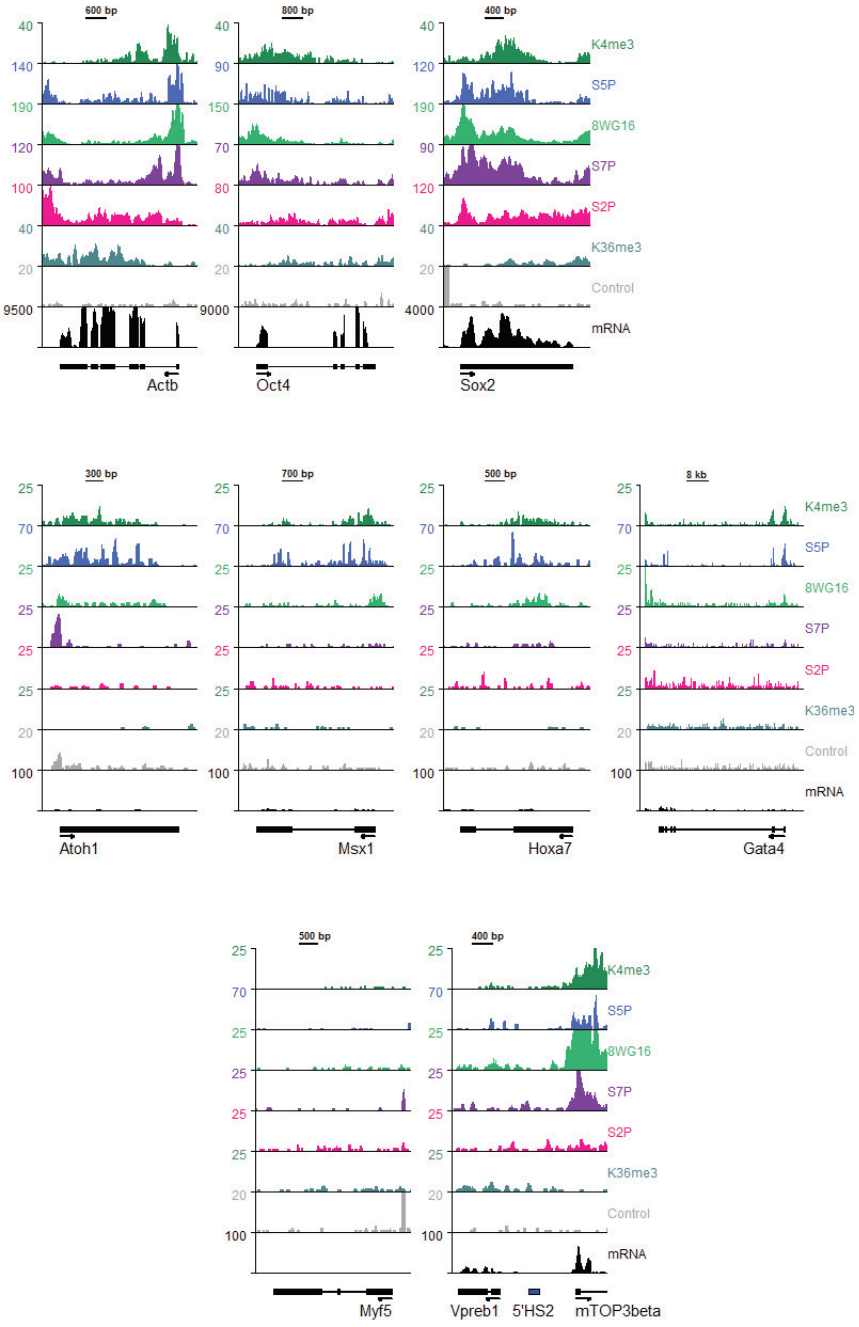


Figure 3.4: ChIP-Seq and mRNA-Seq profiles across single genes. ChIP-Seq profiles are comparable to qPCR, with binding of RNAPII detected at active and PRC-repressed genes but little or low signal at silent genes. Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

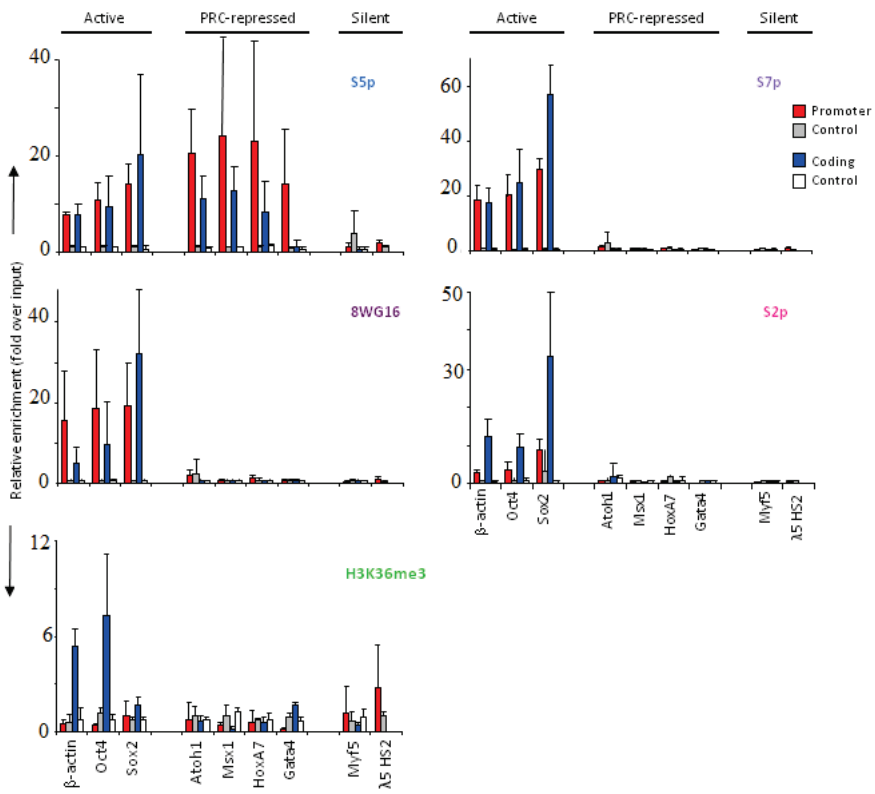


Figure 3.5: Quality control qPCR for RNAPII and H3K36me3 ChIP samples across genes previously studied (Stock et al., 2007) (The same genes depicted in Figure 3.4). Enrichment of S5P, 8WG16, S7P, S2P and H3K36me3 at active, PRC-repressed and silent genes in murine OS25 ES cells was assessed by ChIP and qPCR at promoter (red bars) and coding regions (blue bars), using published primers (Stock et al., 2007; qPCR experiments were performed by E. Brookes). ChIP enrichment is expressed relative to input DNA using the same amount of DNA in the PCR. Background levels (mean enrichment from control antibodies and beads alone) at promoter and coding regions are shown as pale grey or white bars, respectively. Mean and standard deviations are presented from 2-4 independent ChIP experiments.

3.3.4 *RNAPII and transcriptome coverage of the mouse genome in ES cells*

RNAPII is expected to be found at protein-coding and non-coding genes but also at a diverse range of intergenic locations, such as enhancers, or at locations overlapping common enhancer signatures (such as P300 binding and H3K4me1) (Heintzman et al., 2007, Koch et al., 2008, Kim et al., 2010b); RNAPII has also been found to be recruited to insulator regions (Chernukhin et al., 2007), heterochromatin domains in yeast (Steinmetz et al., 2006), or to previously unannotated sites of the human genome (Kim et al., 2005). To reveal the origin of the sequenced reads, we assessed

the number of reads mapping to annotated genes based on the UCSC database (Karolchik et al., 2003), and those mapping to intergenic regions. Mapped reads were classified into the following categories: (i) RefSeq transcripts; (ii) non-RefSeq transcripts (i.e UCSC Known Genes that are not substantiated by a NCBI RefSeq transcript but are supported by at least one GenBank RNA sequence and a Uniprot protein); and (iii) intergenic regions. RefSeq genes were further subdivided into three subregions: 5′ proximal or promoter region (±1kb, upstream and downstream, of the annotated TSS), 3′ proximal region (2kb downstream of the annotated TTS) and the remaining gene body. A summary of the distribution of sequenced reads through the analysed genomic features is given in Figure 3.6 and Table 3.3.

Consistent with previous studies (Kim et al., 2005, Brodsky et al., 2005), RNAPII preferentially associates with genic regions: Between 54% and 60% of the reads from RNAPII libraries overlap with known RefSeq genes; this is significantly different from the percentage of reads from the control library (41%) that overlap the same genomic regions (Chi^2 test p-values < $8x10^{-3}$; for all RNAPII samples when compared to the control). Similarly, H3K36me3 and H3K4me3 are also enriched at gene regions with 64% and 60%, respectively, of reads overlapping known RefSeq genes (Chi^2 test p-values = $3x10^{-6}$ and $1x10^{-4}$, respectively).

Promoter regions are particularly enriched for S5P, 8WG16, S7P and H3K4me3 (Chi^2 test p-values = $4x10^{-7}$, $9x10^{-11}$, $4x10^{-23}$ and $9x10^{-55}$, respectively) and depleted for S2P and H3K36me3 (Chi^2 test p-value = 0.61 and 0.50, respectively), while TTS regions are particularly enriched for S5P, S7P and S2P (Chi^2 test p-value = $8x10^{-3}$, $3x10^{-3}$ and $4x10^{-3}$; Figure 3.6).

For the mRNA-Seq dataset, the vast majority of reads (95%) overlap RefSeq genes, suggesting that these data provide a very deep representation of the transcriptome in mouse ES cells, including detection of genes with very low expression. Finally, between 29% to 39% of RNAPII, H3K36me3 and H3K4me3 reads map to the remaining intergenic regions of the mouse genome, while for mRNA-Seq reads, a sizable minority (2%) of intergenic reads was found. While it is clear that many functional genomic sites may be obscured within the analysed intergenic regions (such as insulator, enhancer or other non-coding sites), much of the observed

enrichment at intergenic regions is low (Figure 3.6) and could potentially be accounted for by unspecific antibody binding and background noise inherent to ChIP data.

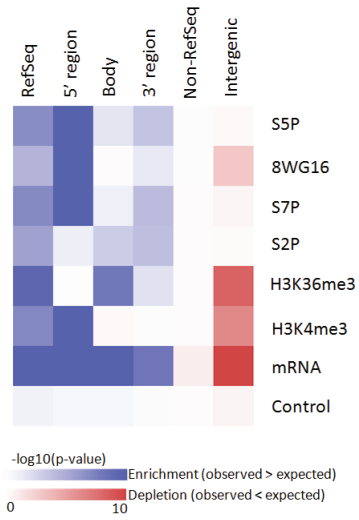


Figure 3.6: Mapped RNAPII and mRNA-Seq reads are strongly biased towards annotated genic regions. Chi-square test p-values were calculated for the observed and expected frequency of reads overlapping each genomic feature. Expected frequency of reads was taken from the % of the genome covered by the genomic feature (values are shown in Table 3.3). The color code represents -log10(p-values) as indicated in the coloured blue (for observed > expected) or red (if observed < expected) scales.

Table 3.3: Summary of read distribution for ChIP-Seq and mRNA-Seq datasets. The distributions were calculated for all aligned reads from each sample. The RefSeq and non-RefSeq set (30844 and 18565 genes, respectively) were derived from UCSC Known Genes dataset as annotated in the UCSC tracks (Karolchik et al., 2003). RefSeq 5′ region was defined as a 2kb region centered at the TSS; 3′ region was defined as 2kb region downstream of the TTS, and "Gene body" refers to the remaining genic region (including introns). UCSC Known Gene coordinates were defined using the genomic start and end coordinates of each transcript as annotated in the UCSC database for the mm9 genome.

		RefSeq					
		5′ region	Body	3′ region	RefSeq (total)	Non-RefSeq	Inter-genic
S5P	Reads (millions)	2.6	9.3	1.3	13.2	1.5	7.4
	Proportion (%)	11.6%	42.1%	6.0%	59.67%	6.7%	33.6%
8WG16	Reads (millions)	4.8	7.5	1.0	13.3	1.7	9.4
	Proportion (%)	19.7%	30.7%	4.1%	54.6%	6.8%	38.6%
S7P	Reads (millions)	1.6	4.4	0.7	6.7	0.8	3.7
	Proportion (%)	14.5%	39.3%	6.4%	60.14%	6.7%	33.2%
S2P	Reads (millions)	1.3	15.7	2.1	19.0	2.2	12.0
	Proportion (%)	3.8%	47.1%	6.2%	57.2%	6.7%	36.1%
H3K36me3	Reads (millions)	0.4	13.5	1.0	14.9	1.5	6.7
	Proportion (%)	1.8%	58.4%	4.5%	64.7%	6.6%	28.7%
H3K4me3	Reads (millions)	2.7	2.7	0.2	5.5	0.6	3.0
	Proportion (%)	29.4%	29.1%	2.0%	60.5%	6.6%	32.9%
mRNA	Reads (millions)	11.1	69.0	8.1	88.1	2.9	1.9
	Proportion (%)	11.9%	74.2%	8.7%	94.89%	3.1%	2.0%
Control	Reads (millions)	0.3	4.3	0.2	4.9	0.8	6.0
	Proportion (%)	3.0%	36.5%	2.1%	41.6%	6.7%	51.7%
Non-redundant size region (Mb)		43.4	842.7	41.6	927.7	193.6	1533.7
% of total genome (coverage)		1.6%	1.6%	31.7%	35.0%	7.3%	57.8%

3.3.5 *RNAPII profiles at highly and minimally expressed genes*

We first investigated typical RNAPII patterns at the extremes of the ES cell transcriptome. To do so, I divided genes into two groups, depending on the expression levels, i.e. top 20% highly expressed (17 to 3156 FPKM; 3772 genes) and bottom 20% least expressed (0 to $8x10^{-3}$ FPKM; 3772 genes).

Average profiles (Figure 3.7) of the most active (20% most expressed genes; "Top20%") genes across the ES cell genome identifies 8WG16 peaking at the TSS and decreasing through coding regions, while S5P peaks at the TSS and extends through coding regions at decreasing levels, showing a small increase again downstream of the TTS. S7P shows enrichment at promoter and coding regions of active genes. The profile for S7P is generally similar to that of S5P at active genes, consistent with Cdk7 catalysing the addition of both modifications (Akhtar et al., 2009, Glover-Cutter et al., 2009). S5P, 8WG16 and S7P are enriched at the region surrounding the TSS, with S5P, 8WG16 and S7P maxima occuring within 40bp of the annotated transcription start site. At active genes, S2P increases throughout coding regions and peaks at approximately 700bp downstream of the 3' end, accompanied by the secondary peak of S5P and S7P. RNAPII accumulation beyond TTS, marked by S2P, S5P and S7P, may reflect transcript 3' end processing, in agreement with the described role for S2P in the recruitment of polyadenylation factors in vivo (Ahn et al., 2004). Inactive genes (20% lowest expression; "Bottom20%") show little or no RNAPII signal, demonstrating that ES cells harbor a large group of genes silent in the absence of RNAPII.

Active chromatin modifications are known to relate to expression and RNAPII state, and this is visible across single genes and average profiles. Our genome-wide chromatin maps reveal a global pattern of H3K36me3 and H3K4me3 in mouse is similar to those previously observed in mouse and human (Mikkelsen et al., 2007, Barski et al., 2007). The level of H3K36me3 enrichment across genes is strongly correlated with the level of S2P and of mRNA which is consistent with a role for S2P in recruiting H3K36me3 HMTs (Krogan et al., 2003, Xiao et al., 2003, Li et al., 2003). H3K36me3 is strongly enriched across transcribed regions of active genes, and is absent from inactive genes. However, the peak in S2P enrichment seen after the 3' ends of genes is not seen in the H3K36me3 profile.

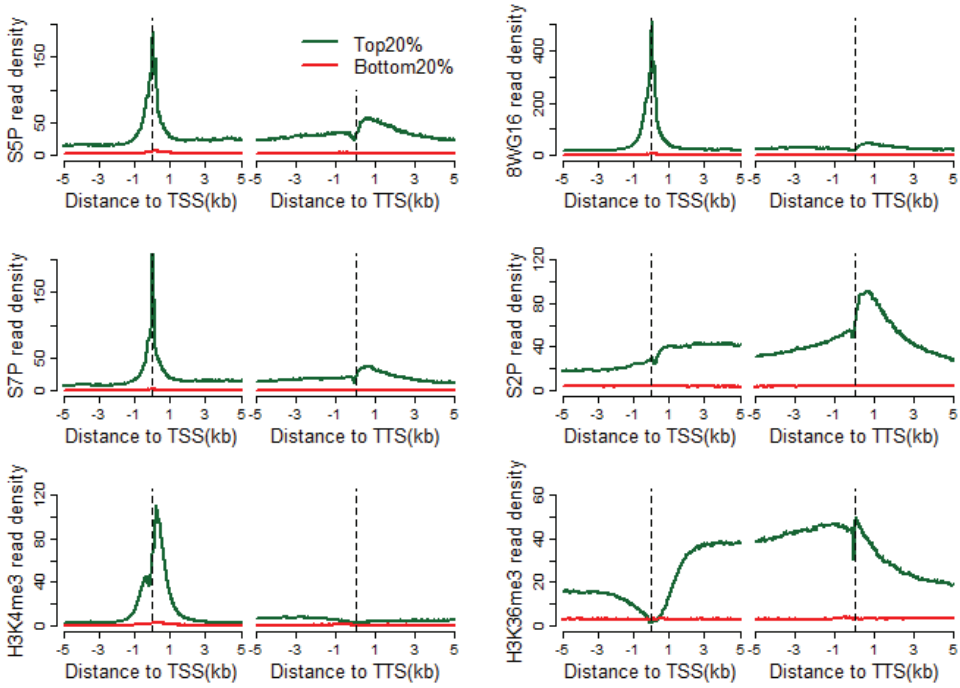


Figure 3.7: Genome-wide ChIP-Seq profiles at active and inactive genes. Average ChIP-Seq read density (depth per 10bp) was computed for the region surrounding the TSS and TTS of highly (Top20%; green) and minimally (Bottom20%; red) expressed genes. Top20% and Bottom20% genes were classified according to FPKM levels. The dashed line indicates the representative position of the TSS and TTS.

In order to quantitatively compare CTD phosphorylation patterns and transcriptional activity, we compared ChIP-Seq signal of RNAPII forms with gene expression levels given in FPKM units across the genome. RNAPII sequencing depth calculated at promoters (for S5P, 8WG16 and S7P) or 3′ region (for S2P) are very well correlated with expression levels (Figure 3.8). S2P, 8WG16 and S7P levels show the higher correlation coefficients. The lower correlation coefficients for S5P promoter levels (Spearman's correlation = 0.76) can be explained by the fact that genes may experience transcription initiation (and thus S5P detection) without being transcribed. The higher correlation of S2P, 8WG16 and S7P levels with gene expression levels suggests that the presence of S2 phosphorylation at genes 3′ end or the presence of S7P or 8WG16 at genes' promoters best describe the expression outcome of a gene.

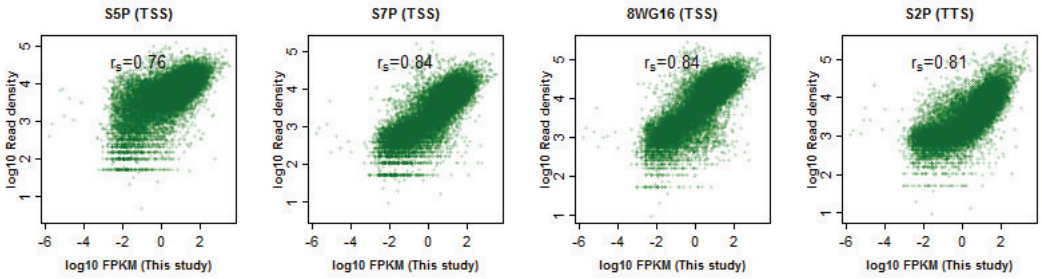


Figure 3.8: RNAPII occupancy predicts mRNA expression. S5P, 8WG16, S7P and S2P density at TSS/TTS windows are highly correlated with mRNA FPKM levels. The best correlation is observed for S2P, S7P and 8WG16 forms. Sequencing depth was calculated in 2kb windows centered at genes TSS (for S5P, 8WG16 and S7P samples) or 2kb downstream of the TTS (for S2P sample). For visualisation purposes, log10 scale is represented. Zeros are omitted in the log10 scale but were taken into account for the calculation of correlation coefficients. All 18860 RefSeq genes were considered. Spearman's rank correlation coefficient is depicted in each plot (r_s).

3.3.6 *RNAPII is associated with both K4 monovalent promoters and K4+K27 bivalent promoters in ES cells*

The first genome-wide characterization of bivalent genes was published by Mikkelsen et al., 2007, where it was stated that bivalent genes are silent in ES cells. Subsequent studies (Stock et al., 2007, Guenther et al., 2007) have used RT-PCR to measure transcript accumulation from bivalent genes and have shown that, while bivalent genes do not produce full-length mRNAs, they undergo transcription initiation or partial elongation.

To investigate the extent of association of bivalent genes with poised RNAPII complexes, we first analysed the average ChIP-seq read distribution of RNAPII phosphorylation and histone modification marks of 2627 bivalent genes classified in Mikkelsen et al., 2007. This analysis identifies a broad enrichment of S5P at bivalent promoters accompanied by H3K4me3. S5P is detected through coding regions but decreases at the TTS of bivalent genes, unlike the increase seen at active genes (Figure 3.9). At the majority of bivalent genes, RNAPII is not marked by S7P or 8WG16, with some exceptions such as *Zic2* and *Isl* which are also associated with S2P and mRNA (see next subsection 3.3.7). Despite abundant detection of S5P, there is little or no detection of S2P or H3K36me3 at the majority of bivalent genes, consistent with the fact that these genes undergo transcription initiation in the

absence of productive elongation (Stock et al., 2007). Lack of S2P marks, known to be present at active genes and to recruit splicing, polyadenylation and H3K36me3 HMTs (Hirose and Ohkuma, 2007), may be crucial in limiting the production of mature full-length mRNAs.

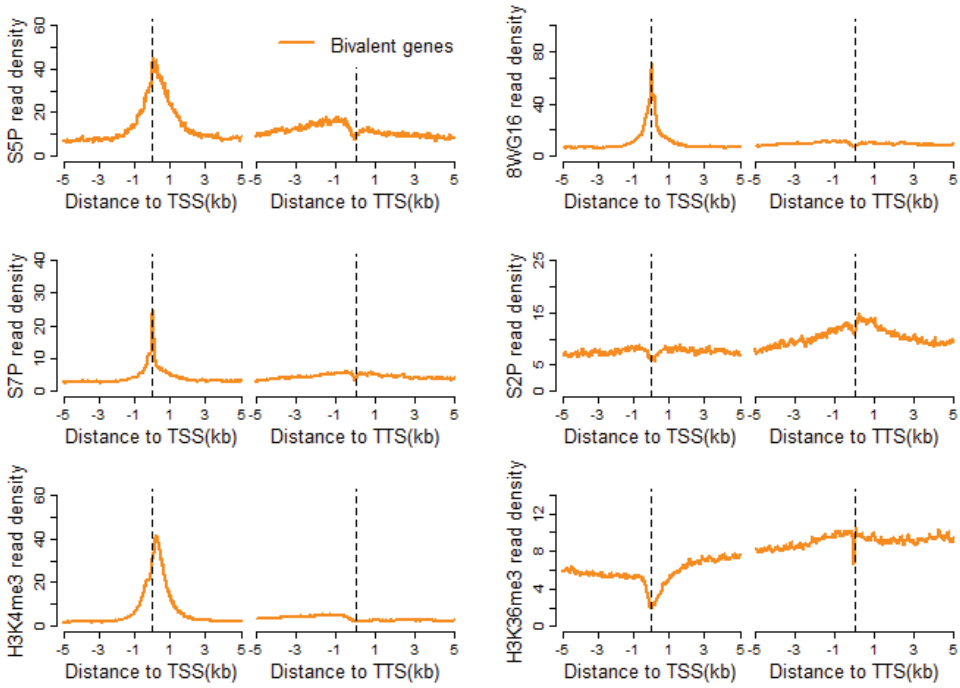


Figure 3.9: Genome-wide ChIP profiles at bivalent genes. Average ChIP-Seq read density (depth per 10bp) was computed for the regions surrounding the TSS and TTS of bivalent genes. Gene list of bivalent gene annotation was downloaded from Mikkelsen et al., 2007. The dashed line indicates the representative position of the TSS and TTS. Note that the Y axis scales are different in relation to Figure 3.7

3.3.7 *RNAPII high-resolution Chip-Seq profiles reveal previously unknown patterns*

While genome-wide maps of RNAPII binding confirm known binding profiles, careful inspection of RNAPII profiles at single genes with distinct expression and epigenetic states reveals a number of interesting cases of unknown combinatorial RNAPII-phosphopatterns (Figure 3.10).

For example, active genes such as *Top2a* and *Eed* show a noncanonical enrichment of S5P at 3'ends and through coding regions. Enrichment of S5P at 3'ends is unexpected according to the current paradigm of the transcriptional cycle where S5P mark is typically known as an "early" transcriptional mark. Another interesting observation

at *Top2a* gene, is the alternative usage of one of its annotated promoters. *Top2a* has two annotated promoters, but only one shows the initiating form of RNAPII (S5P) in ES cells. Our genome-wide catalog of RNAPII phosphorylation and occupancy in combination with mRNA-Seq data contains many other examples that may reveal the alternative usage of many annotated or even novel promoters. Although within the focus of this thesis such genome-wide observations were not yet mined, this is one example of the power of ChIP-Seq analysis.

Other interesting findings relate to the characterization of RNAPII-S7P. Phosphorylation of serine 7 residues has been recently described and linked to the transcription of small nuclear RNAs (Egloff et al., 2007) and protein-coding genes in yeast and human (Chapman et al., 2007, Kim et al., 2010a, Tietjen et al., 2010, Mayer et al., 2010). Human CDK7 kinase, and the yeast homologue Kin28, is responsible for the phosphorylation of S7 and S5 residues of the CTD (Akhtar et al., 2009, Glover-Cutter et al., 2009). In agreement with previous studies, we detect S5P and S7P at promoter regions of active genes. Characterisation of S7P profiles at PRC-repressed genes has not been previously documented. For example, the PRC-repressed gene *Nr2e1* shows prominent S5P promoter peaks while detection of S7P is low or absent (Figure 3.10), suggesting that specific S5P kinases may act or that S7 residues are not accessible for phosphorylation due to the existence of other CTD modifications. In contrast, other known PRC-repressed genes such as *Isl1* and *Zic2* (classified as "bivalent" in Mikkelsen et al. [2007]) show S7P enrichment, either more localized at the promoter (*Isl1*) or widespread along the gene body (*Zic2*). This pattern is accompanied by S5P and 8WG16 enrichments (Figure 3.10). At *Zic2*, this "active-like" RNAPII profile (S5P+8WG16+S7P+) is also accompanied by clear S2P enrichment across *Zic2* gene body and mRNA levels which are clearly visible over the exons of *Zic2*. The genomic or expression constraints that dictate the distinct RNAPII conformations at different groups of PRC-repressed genes merit further investigation and these contrasting patterns at subclasses of PRC-repressed genes will be further investigated in Chapter 5.

Furthermore, RNAPII binding profiles seem to reflect other interesting transcriptional events, such as transcription of non-coding RNA genes (Figure 3.10).

For example, *Kcnq1ot1*, which is a ncRNA known to play a role in the silencing of genes in the *Kcnq1* domain (Thakur et al., 2004), is clearly transcribed by RNAPII and expressed in ES cells. Another example of an expressed ncRNA is *Malat1*, a highly conserved lncRNAs in mammals well known for its overexpression in many cancers (Tripathi et al., 2010).

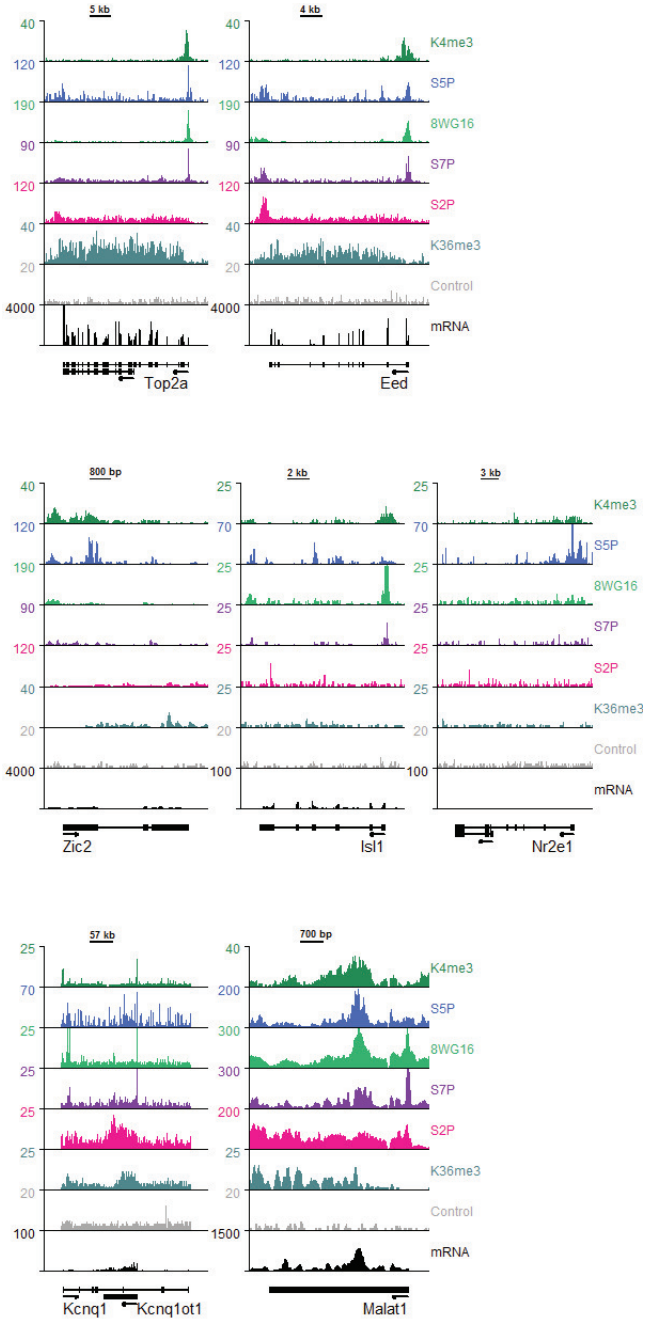


Figure 3.10: Occupancy of RNAPII, H3K36me3 and H3K4me3 at single genes. ChIP-Seq and mRNA-Seq profiles at representative active genes (*Top2a*, *Eed*), genes classified as bivalent (Mikkelsen et al., 2007; *Nr2e1*, *Isl1* and *Zic2*) and non-coding transcripts (*Kcnq1ot1*, *Malat1*). Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

3.3.8 *Comparison of RNAPII ChIP-Seq datasets with published datasets*

ChIP-Seq datasets for S5P and S2P in ES cells were recently published (Rahl et al., 2010). In these publicly available ChIP-seq datasets (Rahl et al., 2010), S5P was detected with a different antibody (ab513111) and S2P with the same antibody (H5) used herein. Published datasets were downloaded from GEO database as fastq files (GSE20530). Reads were re-mapped to mm9 genome assembly using Bowtie aligner (Langmead et al., 2009) as described in Chapter 2. A summary of the number of reads mapped to the mouse genome is presented in Table 3.4. Output alignment files were subsequently processed using python scripts to generate genome-wide average profiles at "Top20%", bivalent, and "Bottom20%" genes.

Table 3.4: List of published RNAPII ChIP-Seq datasets used in this study.

ChIP-Seq dataset	Antibody	Nr of runs	Mapped reads (millions)
S5P (Rahl et al., 2010)	ab5131 (Abcam) (GSM515662)	1	3.6
S2P (Rahl et al., 2010)	H5 (ab24758, Abcam) (GSM515663)	3	10.7

Published datasets could detect some features of RNAPII modification at active genes (such as S5P peaks at TSSs), but missed others (S5P enrichment at TTS; Figure 3.11). In the published datasets, at 20% most active genes S5P is not detected throughout coding regions or at the TTS, in contrast to what we observe (Figure 3.11a). This can be explained by steric hindrance of additional CTD modifications to ab5131 binding. For example, ELISA of ab5131 and 4H8 shows that 4H8, but not ab5131, can detect S5P when S2 is phosphorylated in the same CTD repeat (Brookes, de Santiago et al., 2011; submitted). Therefore, doubly phosphorylated S2P-S5P repeats of the CTD may predominate during elongation.

There are also significant differences in the S2P profiles at active genes between the datasets, although the same antibody (H5) was used. Firstly, our dataset have higher depth (compare S2P in Figure 3.7 and Figure 3.11). Secondly, the average profile for the Top20% expressed genes, in the published S2P dataset (Rahl et al., 2010), shows a

promoter peak which is not seen in our dataset, and moreover not mentioned in the original study (Rahl et al., 2010; Figure 3.11a). As the antibody clone used in both studies is the same (H5), the unexpected S2P peak at the TSS from the published data may result from experimental differences (we fix cells directly in media, instead of in PBS, and use phosphatase inhibitors throughout the ChIP procedure to avoid interfering with RNAPII processivity during washes in PBS or RNAPII phosphorylation) or from antibody batch variation (not observed between different batches used for ChIP in our laboratory for the last six years). Single gene profiles confirm the presence of S2P peaks at the 5′ end of active genes from the published datasets contrary to what we find in our own dataset (Figure 3.11b); the data from Rahl et al. [2010] disagrees with the current paradigm of the transcription cycle, where S2P increases during transcript elongation and peaks at the 3′ end. Nonetheless, despite the differences observed, a good correlation is observed when comparing the depth of coverage at 2kb windows, either centered at TSS (for the S5P comparison: Spearman's correlation coefficient = 0.85; not shown) or downstream of the TTS (for the S2P comparison; Spearman's correlation coefficient = 0.82; not shown).

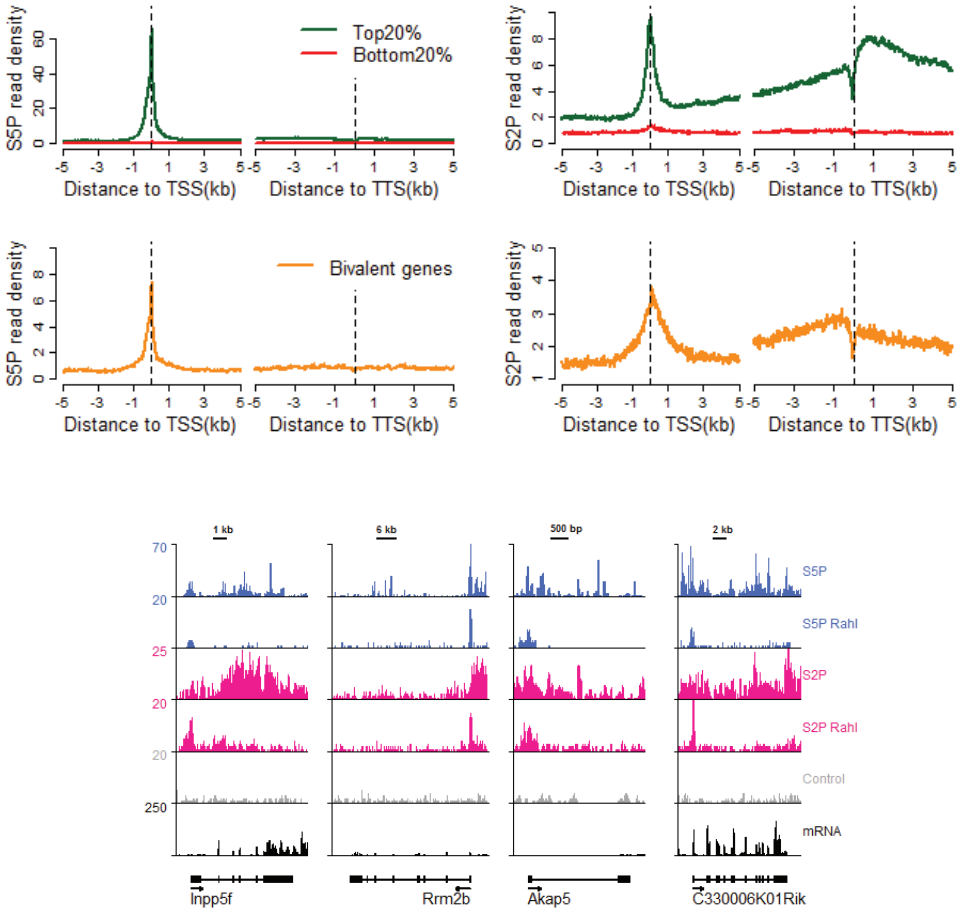


Figure 3.11: Comparison of newly generated and published ChIP-Seq datasets for S5P and S2P forms of the RNAPII. (a) Rahl et al. [2010] RNAPII ChIP-Seq profiles at 20% most expressed (Top20%), least expressed (Bottom20%) genes, and bivalent genes. Average ChIP-Seq read density (depth per 10bp) was computed from publicly available S5P and S2P datasets (Rahl et al., 2010). The dashed line indicates the representative position of the TSS and TTS. (b) Single gene profiles demonstrate the presence of an unexpected peak of S2P at active gene promoters detected when using data from Rahl et al., 2010 (compare "S2P" with "S2P-Rahl"). Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

3.3.9 *Events associated with transcriptional elongation occur at a proportion of bivalent genes*

In Stock et al., 2007 we have demonstrated that for a panel of bivalent genes (*Atoh1*, *Nkx2-2*, *Msx1*, *Nkx2-9*, *Cdx2*, *Gata4*) RNAPII produces low levels of transcripts that are sensitive to α-amanitin, and transcripts undergo splicing of at least the first intron,

suggesting that S5 phosphorylation, in the absence of detectable S2 phosphorylation, is congruent with extensive elongation in this context (Stock et al., 2007).

I began exploring the transcriptional activity of bivalent genes by analyzing and comparing several independent microarray transcriptome studies (Mikkelsen et al., 2007, Sene et al., 2007), two genome-wide mRNA-Seq transcriptome studies (this study, and Cloonan et al., 2008) and two proteome studies in mouse ES cells (Graumann et al., 2008, Van Hoof et al., 2006).

Two microarray studies from independent laboratories (Mikkelsen et al., 2007, Sene et al., 2007) measured global gene expression in mouse V6.5 ES cells using Affymetrix GeneChip expression arrays made of 3' probes; the mRNA-Seq study from Cloonan et al., 2008 used ABI SOLiD technology, and finally our novel mRNA-Seq dataset was obtained using Illumina technology. In all transcriptome studies, cDNA was generated from poly(A)+ mRNA transcripts. Table 3.5 contains details about differences between the considered studies.

Table 3.5: List of mRNA datasets used.

Expression dataset	Protocol	Platform	Cell Type	Nr of mapped reads
This study	PolyA selection, ribosome depletion	Illumina (mRNA-Seq)	ES-OS25	93 million
Cloonan et al., 2008 (SRA000306)	PolyA selection, ribosome depletion	SOLiD (mRNA-Seq)	ES-W9.5	8.4 million
Mikkelsen et al., 2007	PolyA selection	Affymetrix GeneChip Mouse Genome 430 2.0	ES-V6.5	
Sene et al., 2007	PolyA selection	Affymetrix GeneChip Mouse Genome 430A and 430B	ES-V6.5	

Using our and Cloonan et al. [2008] sequenced reads, we estimated the expression levels of genes. Expression levels were measured in ''fragments per kilobase of exon model per million mapped reads'' (FPKM) (Trapnell et al., 2009). We next analysed

all genes with expression levels greater then zero (FPKM > 0). Using this criterion, 15653 and 14218 expressed genes (out of a total of 18860 UCSC genes) were detected, according to our own and Cloonan et al. [2008] mRNA-Seq studies, respectively. From published gene lists, a total of 12984 expressed genes were retrieved from Mikkelsen et al. [2007] dataset, while Sene et al. [2007] dataset contained expression levels for only 7876 genes. There are many challenges in re-analysing microarray data sets and comparing between mRNA-Seq and microarray studies, such as variations in experimental design, including variability in microarray or sequencing platforms, sample preparation, and data processing (Edwards et al., 2008). Other challenges arise due to the lack of supplementary information or ambiguous information (e.g. ambiguous gene IDs). To compare the results for the set of common RefSeq genes quantitatively, I correlated gene expression levels between all four datasets (Figure 3.12). A pairwise correlation of gene expression levels shows that microarray data sets are moderately correlated (Spearman's correlation coefficient = 0.5; p-value $< 10^{-16}$). Sene et al. [2007] data is also only moderately correlated with our own study and Cloonan et al. [2008] (Spearman's correlation coefficient = 0.6 and 0.5; respectively; p-value $< 10^{-16}$). The fact that Sene et al. [2007] has the lowest number of genes could account for the lower correlations observed. Mikkelsen et al. [2007], Cloonan et al. [2008] and our study are well correlated (Spearman's correlation ranging from 0.78 to 0.86); the correlation coefficient is higher between the two mRNA-Seq datasets, as expected. Although, a good correlation is observed between Mikkelsen et al. [2007] and both mRNA-Seq studies (Cloonan et al., 2008 and our own), the correlation is mostly lost for low expressed genes. In Figure 3.12 it is clear that Mikkelsen et al. [2007] has a large cohort of genes with very low expression, which probably result from the insensitivity of microarrays to separate "signal" from "noise" at lower expression levels.

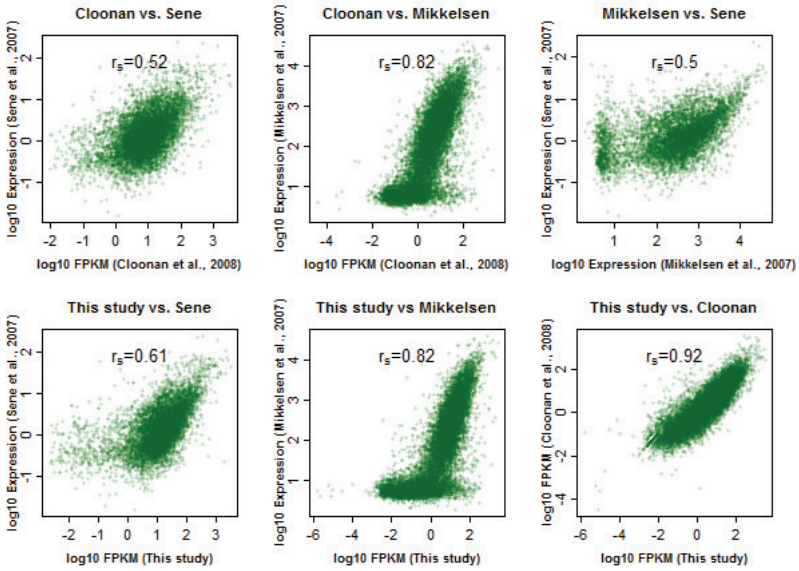


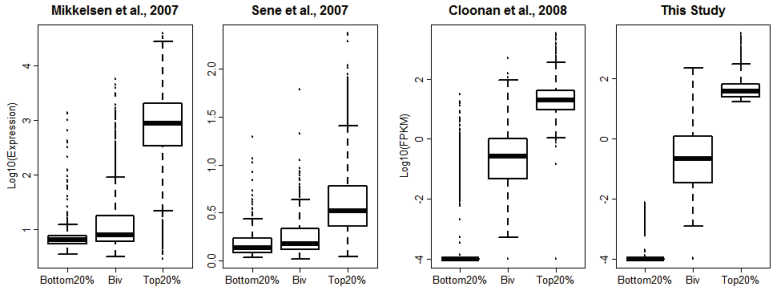
Figure 3.12: Comparison of expression levels between independent mRNA studies. sequencing lanes of the mRNA library. For visualisation purposes, expression levels are represented in log10 scale. Zeros are omitted in the log10 scale but were taken into account for the calculation of correlation coefficients. Spearman's rank correlation coefficient is depicted in each scatter plot (r_s).

Given the possibility to detect low expressed transcripts from mRNA-seq data, I next examined the expression levels of genes marked by bivalent domains classified according to Mikkelsen et al. [2007]. According to Sene et al. [2007] and Mikkelsen et al. [2007] datasets, bivalent genes have low expression, with the overall distribution being similar to that for Bottom20% genes. However, according to both mRNA-Seq studies, a larger range of expression levels at bivalent genes is observed, with some genes showing intermediate and high expression (Figure 3.13a). In total, 402 expressed bivalent genes were consistently detected in all gene lists (Figure 3.13b).

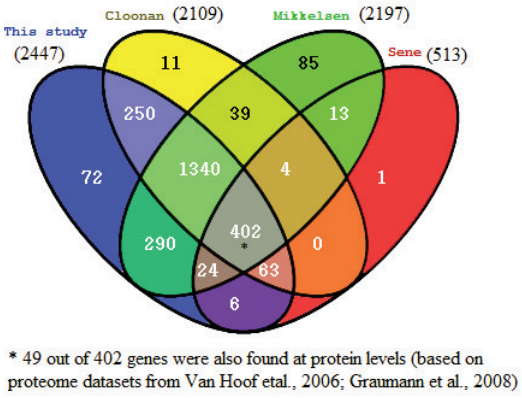
To investigate whether transcripts from bivalent genes are detected at the protein level, I used two published proteome datasets obtained by a mass spectrometry approach in mouse D3 ES cells (Van Hoof et al., 2006) and SILAC methodology in mouse R1 ES cells (Graumann et al., 2008). A total of 111 genes were identified by one or both proteome studies, and out of these 111 genes, 49 were also detected as expressed consistently in all four analysed studies.

Bivalent genes have been long known to be enriched for developmental regulator genes (Mikkelsen et al., 2007, Ku et al., 2008). Gene Ontology (GO) analysis revealed

that 402 expressed bivalent genes contains several genes encoding TFs (*Hoxa1*, *Pax8*, *Six1*, *Igfbp3*, *Id3*) and are enriched for processes relating to "multicellular organismal development" (p-value = 7.9×10^{-14}; Fisher exact test, bonferroni correction), "regulation of developmental process" (p-value = 9.1×10^{-5}), and "organ morphogenesis" (p-value = 7.3×10^{-6}). Despite the over-representation of developmental regulator genes, the analysed set also includes genes encoding membrane proteins, as processes related to "cell adhesion" (p-value = 4.2×10^{-4}) and "cell motion" (p-value = 3.3×10^{-3}) were also enriched.



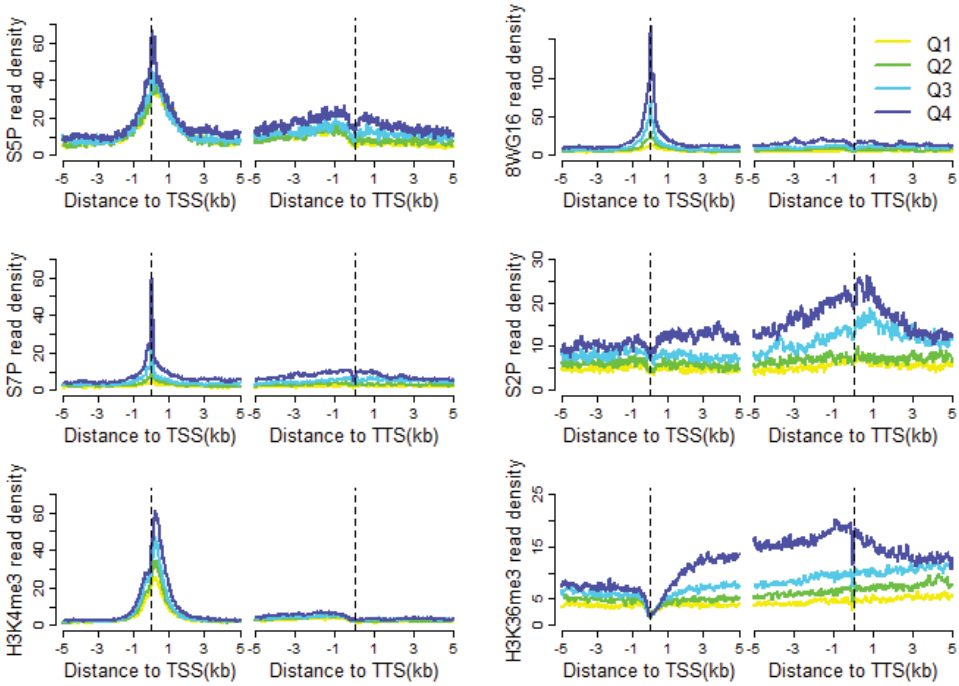
(a)



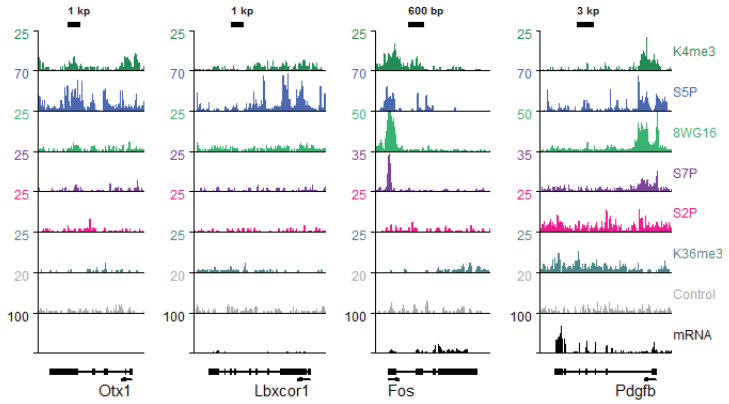
(b)

Figure 3.13: Expression of bivalent genes is detected in gene expression studies. (a) Comparison of expression levels of bivalent (Biv), highly (Top20%) and least (Bottom20%) expressed genes. Bivalent genes were classified in Mikkelsen et al. [2007]. Top20% and Bottom20% genes were classified according to FPKM levels calculated in this study. For visualisation purposes, expression levels are represented in log10 scale. To avoid taking the log of zero, a pseudo value of 0.0001 (in the case of FPKM) was added prior taking the log. GeneChip microarrays detect lower mRNA levels for bivalent genes then mRNA-Seq studies. (b) Venn diagram showing the overlap between bivalent genes classified as expressed in each indicated gene list. Expression studies differ in experimental conditions, cell lines and analytical steps (see Table 3.5 for details). Mikkelsen et al., 2007 and Sene et al., 2007 used GeneChip microarrays, while Cloonan et al., 2008 and our own study used mRNA-Seq. Numbers in parenthesis represent the total number of bivalent genes found.

Next, I classified bivalent genes into groups based on their FPKM values: bottom 25th percentile (Q1; FPKM <= 0.04), middle 25th quartiles (Q2 and Q3; 0.04 < FPKM $\leq$ 0.22 and 0.22 < FPKM $\leq$ 1.18, respectively), and top 25th quartile (Q4; FPKM > 1.18). Average profiles representing RNAPII modification at four quantiles of bivalent genes defined expression levels show that expression is positively associated with presence of 8WG16, S7P and S2P (Figure 3.14a). Inspection of ChIP-Seq read distributions at single bivalent genes representative of each expression quartile (Figure 3.14b) confirmed distinct RNAPII profiles, with silent bivalent genes being generally associated with S5P only, whereas actively-expressed bivalent genes were occupied by both S5P and S2P.



(a)



(b)

Figure 3.14: Expression of bivalent genes is positively associated with presence of 8WG16, S7P, S2P and H3K36me3. (a) Genome-wide ChIP profiles at bivalent genes classified into groups based on their FPKM values: bottom 25th percentile (Q1; FPKM $\leq$ 0.04), middle 25th quartiles (Q2 and Q3; 0.04 < FPKM $\leq$ 0.22 and 0.22 < FPKM $\leq$ 1.18), and top 25th quartile (Q4; FPKM > 1.18). Average ChIP-Seq read density (depth per 10bp) was computed for the region surrounding the TSS and TTS of the four classes of bivalent genes. Expression levels of bivalent genes are positively associated with presence of 8WG16, S7P, S2P and H3K36me3. (b) ChIP-Seq and mRNA-Seq profiles at bivalent genes representative of each expression quartile: *Otx1* (Q1, FPKM = 0.02), *Lbxcor1* (Q2; FPKM = 0.15), *Fos* (Q3; FPKM = 0.73) and *Pdgfb* (Q4; FPKM = 3.49). Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

3.4 DISCUSSION

In this chapter, ChIP-Seq technology was used to map RNAPII binding genome-wide and investigate the CTD code in mouse ES cells. Our results reveal that the configuration of CTD-phosphorylation differs among genes with different expression and epigenetic states. We examined qPCR-ChIP, ChIP-Seq and mRNA-Seq profiles across genes for which the pattern of RNAPII phosphorylation has been well documented as well as other newly examined genes where new combinatorial RNAPII-phosphopatterns was revealed. This study provides a genome-wide catalog of CTD modifications in RNAPII with deep coverage of the whole genome. The data reveals (i) the existence of non-canonical patterns of RNAPII CTD modifications at active and bivalent genes; (ii) the existence of varying patterns of RNAPII within bivalent genes; (iii) the co-occurrence of PRC repression and active RNAPII at a subclass of bivalent genes. Taken together, our data shows that RNAPII regulation is more complex then previously thought.

3.4.1 *RNAPII configuration at active and inactive genes*

Analysis of the top 20% most active genes (3772 genes; Figure 3.7) detects peaks of RNAPII-S5P, -8WG16 and -S7P and H3K4me3 at promoter regions and detection of S5P, and increasing levels of S2P and H3K36me3 through coding regions. Downstream of 3' ends we identify a region of S2P enrichment (peaking at approximately 700bp downstream the annotated TTS), alongside with smaller but detectable increased levels of S7P and S5P. The preferential enrichment of S5P, S7P and 8WG16 at promoters and S2P and H3K36me3 at gene's coding and TTS regions was also shown by the distribution of reads into gene regions of interest (TSS, gene body and TTS; Figure 3.6). We have shown that S2P is a good indicator of mRNA production, consistent with the described role for S2P in transcript elongation, including the recruitment of splicing and polyadenylation factors required for mRNA maturation (Hirose and Ohkuma, 2007).

The more recently described RNAPII mark, S7P, was also studied here. The known function of S7P is associated with the recruitment of the Integrator complex to

snRNA-encoding genes (Egloff et al., 2007). The patterns of S7P marks observed here by ChIP-Seq across transcribed regions and at the TTS, suggest that S7P may also contribute to efficient mRNA processing and elongation of RNAPII at protein-coding genes. Similarly to S5P, S7P is enriched at 5' ends of transcribed genes (Figure 3.4, Figure 3.7). As RNAPII progresses through genes, both marks exhibit a drop but still high levels are retained across gene bodies, suggesting that a combination of S2P, S5P and S7P may contribute to efficient elongation and termination. Consistent with these results, *in vitro* studies have shown that the HMT that methylates H3K36, Set2, preferentially recognise a doubly marked CTD (S5P and S2P), rather then S2P alone (Kizer et al., 2005). Furthermore, S5P and S7P enrichment at the gene body and TTS is consistent with the recently described role of Cdk7 in regulating elongation and termination (Glover-Cutter et al., 2009), as Cdk7 has been described as the kinase responsible for the phosphorylation of both S5 and S7 CTD residues (Akhtar et al., 2009, Glover-Cutter et al., 2009). The association of both S5P and S7P to gene's bodies and TTSs of expressed genes has not yet been discussed in current genome-wide datasets, including the one analysed here (Rahl et al., 2010) and other genome-wide maps of the RNAPII CTD modifications in yeast (Kim et al., 2010a, Mayer et al., 2010, Tietjen et al., 2010), thus demonstrating that this data provides a very deep representation of the RNAPII CTD code.

The RNAPII enrichment identified at the TTS is characterised by slight increases in the enrichment of S5P, S7P and a more prominent enrichment of S2P immediately downstream of the TTS of actively transcribed genes. The RNAPII peak downstream of the TTS may reflect slower kinetics of RNAPII in the region where polyadenylation and termination signals take place and this paused site is probably required for correct processing of mRNA.

The profiles of H3K4me3 overlaps with the profile of S5P only to some degree. Whereas high levels of S5P are retained as RNAPII travels along genes and it increases at the TTS, H3K4me3 signal is not detected at the TTS of highly transcribed genes (Figure 3.4, Figure 3.7, Figure 3.10), suggesting that the interaction between S5P and H3K4 HMTs is stronger at the sight of transcription initiation and may be disrupted by S2P. During elongation, the S2P CTD associates with H3K36-specific

HMTs (Krogan et al., 2003, Xiao et al., 2003, Li et al., 2003), and is thus in agreement with what is expected, both S2P and H3K36me3 are highly enriched throughout the gene body of active genes. However, while S2P peaks downstream of the TTS, H3K36me3 shows a slight drop downstream of the 3' end, suggesting that the interaction between S2P and H3K36me3 HMTs is disrupted which might be crucial for the regulation of transcription termination.

Aggregate profiles across the bottom 20% least expressed genes (3772 genes) show no enrichment of RNAPII signal throughout the whole transcribed region. In Chapter 4, we show that this large proportion of silenced genes is not silent through mechanisms dependent on PRC-repression as they were depleted of PRC proteins (Ezh2, Suz12 and Ring1B) and the two chromatin modification marks associated with PRC-mediated gene repression (H3K27me3 and H2Aub1). It is possible that these genes are silenced through other mechanisms independent of PRC-repression, such as H3K9me3 and H4K20me3 or DNA methylation, or that these genes are passively repressed and that their silent state is simply due to the lack of activating factors. Investigating other silencing factors at this group of genes may further elucidate their mechanisms of silencing.

3.4.2 *PRC-repressed RNAPII has an unusual combination of posttranslational modifications in ES cells*

In ES cells, H3K27me3 can co-exist with H3K4me3 in many silent developmental regulator genes (Azuara et al., 2006, Bernstein et al., 2006). Although genes that carry this 'bivalent' modification are repressed, they may also carry appreciable amounts of RNAPII (Stock et al., 2007). Inspection of RNAPII CTD modifications genome-wide (Figure 3.9, Figure 3.10) showed that the majority of bivalent genes do not display S7P or 8WG16 at their promoters, and neither the active form of RNAPII (S2P) or the elongation marker H3K36me3, but S5P is detected at the genes promoter and throughoutt transcribed regions. To our knowledge, the progress of RNAPII-S5P in the absence of S2 phosphorylation was not thought to occur (Buratowski, 2009). The absence of S2P from bivalent chromatin is reinforced by lack of H3K36me3. Surprisingly, binding of 8WG16 (an antibody that recognizes non-phosphorylated S2

residues) to bivalent promoters is also low or undetectable. The lack of recognition of RNAPII complexes at bivalent promoters by antibodies that recognize the phosphorylated (H5) or non-phosphorylated (8WG16) S2 residues suggests that these residues might be occluded, either due to another CTD modification, binding of an external factor, or due to a CTD conformation that is inaccessible to these antibodies. While one could argue that other undetected modifications, such as S7P, might also be invisible to antibody detection, S5P residues are not occluded from antibody binding.

3.4.3 *Transcription of bivalent genes*

In this study I have used ChIP-seq to generate genome-wide maps of several RNAPII CTD modifications and two histone modification marks associated with gene transcription initiation (H3K4me3) and elongation (H3K36me3). All bivalent genes show an enrichment of H3K4me3 and RNAPII-S5P at their promoters, nonetheless we have identified distinct types of enrichment profiles of S5P, S7P, 8WG16 and S2P at bivalent genes (Figure 3.10). First, the vast majority of bivalent genes show an enrichment of Ser5P-RNAPII at their promoters and coding regions in the absence of S2P and H3K36me3, suggesting extensive transcription, while the absence of S2P implies incompatibility with efficient RNA processing (Stock et al., 2007). Second, and surprisingly, full length transcripts were detected for some bivalent genes (Figure 3.13). This finding was consistant with the existence of active RNAPII marks (Figure 3.14), and, for a few bivalent genes, with detection of expression at the protein level. We hypothesize that the previous suggestion that bivalent genes are repressed in pluripotent ES cells is due to the insensitivity to low mRNA signals on microarray analysis. Taking advantage of high coverage RNA-Seq and ChIP-Seq data, our data suggests that in ES cells bivalent genes are transcribed producing either a variety of short, 5'-abortive transcripts (Guenther et al., 2007, Stock et al., 2007) or full length transcripts.

3.4.3.1 *Pervasive transcription from silent bivalent genes*

There have been several studies reporting general widespread transcription from ES cells (Turner, 2008, Efroni et al., 2008). Similarly, several lines of evidence support the view that low level transcription occurs at bivalent genes in ES cells. For instance, we have shown in a previous publication that low levels of α-amanitin-sensitive transcripts are produced at a panel of 9 analysed bivalent genes (Stock et al., 2007); splicing of their first intron was detected, but it is not known if transcription extends through the entire coding region. Other studies have shown that bivalent genes are transcribed and produce a variety of short, 5'-abortive transcripts (Kanhere et al., 2010, Guenther et al., 2007). Furthermore, the majority of bivalent promoters are free of DNA methylation (Meissner et al., 2008, Fouse et al., 2008) and thus reside in more dynamic chromatin environment. The active environment could create an accessible chromatin state, more permissive to transcription (Mattout and Meshorer, 2010), and thus could reflect less stringent control of transcription initiation at bivalent genes.

The functional relevance of bivalent gene transcription remains obscure. On one hand, global transcription in ES cells could simply be a consequence of their accessible chromatin structure with no specific consequence in itself. In this case, non-functional transcripts would be continually degraded. On the other hand, transcription itself could play a role in gene regulation. Interestingly, transcriptome research in primary T cells and ES cells revealed the ncRNAs can interact with PRC2 and are thought to play a role in the association of PRC2 with its target gene (Kanhere et al., 2010). Alternatively, pervasive transcription at bivalent promoters could also have a global influence in their chromatin context. For instance, it is known that H3K4 trimethylation and DNA methylation are mutually exclusive (Weber et al., 2007) and thus transcription could have an active role in maintaining these genes protected from DNA methylation and thus helping to preserve the bivalent state, compatible with more efficient gene activation. It will be interesting to investigate the mechanisms by which transcripts from bivalent genes are produced, without being fully processed, and their possible roles in embryonic development. A study in *Drosophila* reported the existence of alternative TSSs that precede long introns ('macrointrons') in developmental genes (Manak et al., 2006). As suggested by

the authors, there seemed to be no reason for this particular combination other than transcription itself, or to communicate with distant regions where transcription is taking place and participate in their regulation. Since a great amount of non-annotated transcription is observed at bivalent genes and considering their bias for larger intron lengths, it will be interesting to learn whether similar global transcription mechanisms are used in bivalent genes by vertebrates.

3.4.3.2 *Bivalent genes associated with active transcription*

I have shown that low levels of mRNA from bivalent genes can be detected by microarray (Mikkelsen et al., 2007, Sene et al., 2007) or RNA-seq (this study, and Cloonan et al., 2008) analysis, suggesting that transcript maturation may occur to a limited extent. In comparison to microarrays, mRNA-Seq data show more clearly that a given proportion of bivalent genes are expressed, this could be explained by insensitivity of microarrays in detecting low levels of transcription (Hebenstreit and Teichmann, 2011).

The analysis presented here shows a positive association between RNA expression levels and the levels of "active" RNAPII marks. The different patterns of CTD phosphorylation, alongside different H3K4me3 and H3K36me3 enrichments, seem to reflect different levels of transcriptional activity at bivalent genes. In particular, highly expressed bivalent genes are enriched for S7P, 8WG16 and S2P in contrast to "silent" bivalent genes, for which mRNA is not detectable. Therefore, in contrast to other more restrictive forms of silencing such as DNA methylation, Polycomb repression does not necessarily prevent sporadic full-length gene transcription. This hypothesis is consistent with a previous study in human ES cells in which sporadic expression of bivalent genes was detected at the single cell level (Gibson et al., 2009).

In Chapter 5 I further explore the possibility that the detection of such "active" bivalent genes might be a simple read-out of the presence of two distinct subpopulations of cells in which the genes are either active or silent.

4

POLYCOMB REPRESSION IN THE PRESENCE OF RNAPII

4.1 RESEARCH MOTIVATION

In ES pluripotent cells, PRCs are thought to silence important developmental regulator and lineage-specific genes while keeping them 'poised' for future activation during differentiation (Azuara et al., 2006, Bernstein et al., 2006). In Chapter 3 I describe different RNAPII phosphorylation patterns found at different bivalent genes which suggested that different levels of PRCs, active (H3K4me3) and silent (H3K27me3) epigenetic marks, and RNAPII can vary among bivalent genes. My aim in this project was to further explore RNAPII occupancy at PRC-repressed genes genome-wide and to investigate if specific RNAPII phosphorylation patterns characterise different classes of PRC-repressed genes. I aimed to perform integrative bioinformatic analysis of ChIP-Seq data of RNAPII modifications, PRC proteins and PRC-mediated chromatin modifications and mRNA-Seq data. To investigate the combinatorial association between RNAPII and PRCs, I also aimed to develop a new classification scheme to classify genes according to the presence/absence of RNAPII or histone modification ChIP-Seq signal.

The experimental work presented in this chapter was carried out by E. Brookes.

4.2 RESULTS

4.2.1 *Genome-wide mapping of Polycomb and histone modification data in ES cells*

To investigate the RNAPII occupancy and configuration at PRC-repressed genes, we performed ChIP-seq for the PRC1 catalytic subunit Ring1B, together with the histone modification that it catalyzes, H2Aub1, and utilised publicly available ChIP-seq data for PRC2 (Ezh2, Suz12) (Ku et al., 2008) and H3K27me3 (Mikkelsen et al., 2007) and, for comparison, a Ring1B ChIP-Seq dataset (Ku et al., 2008). Ring1B and H2Aub1 ChIP samples for wild-type mouse OS25 ES cells were sequenced using Illumina's Genome Analyzer II in two to three runs. Sequenced reads were subsequently aligned to the mm9 assembly of the UCSC mouse genome (Karolchik et al., 2003) using the Illumina alignment software Eland as described in Chapter 2. Raw sequenced reads from published datasets (H3K27me3, Ezh2, Suz12 and Ring1B) were retrieved from GEO (GSM307619, GSM327668, GSM327667 and GSM327669, respectively) as fastq files and remapped to the mm9 genome using Bowtie alignment software (Langmead et al., 2009) as described in Chapter 2. A summary of the number of DNA reads sequenced at each run, and their mapping to the mouse genome is presented in Table 4.1.

Prior to ChIP-seq library generation, we performed qPCR at a panel of well-characterized active, PRC-repressed and silent genes (Stock et al., 2007) to confirm the quality of ChIP samples Figure 4.1. At active (*ß-actin, Oct4, Sox2*) and silent genes (*Myf5*) or genomic regions (5'HS2), little or no enrichment for Polycomb subunits (Ring1B, Ezh2, Suz12) or associated histone modifications (H2Aub1 and H3K27me3) is detected (Figure 4.2), while high levels are detected at PRC-repressed genes (*Atoh1*, *Msx1*, *HoxA7*, *Gata4*). ChIP-Seq profiles are comparable to qPCR results for the two datasets generated herein (Ring1B ad H2Aub1) confirming the presence of PRC function at promoters and coding regions of the selected PRC-repressed genes (Stock et al., 2007).

Table 4.1: List of ChIP-Seq datasets used in this study. The percentage of mapped reads corresponds to the number of reads that map to the mm9 genome out of the total number of reads that pass Illumina's quality chastity filter.

ChIP-Seq dataset	Antibody	Nr of Runs	Mapped reads (millions)	% of mapped reads (Illumina alignment)
H2Aub1	E6C5 (Upstate)	2	lane 1: 7.0 lane 2: 11.1	lane 1: 72.1% lane 2: 72.9%
Ring1B	A-20 (Atsuta et al., 2001)	3	lane 1: 4.3 lane 2: 4.4 lane 3: 18.9	lane 1: 73.0% lane 2: 73.1% lane 3: 73.9%
Ring1B (Ku et al., 2008)	A-20 (Atsuta et al., 2001)	2	4.2	
Ezh2	39103(Active Motif)	2	9.5	
Suz12	ab12073 (Abcam)	3	13.3	
H3K27me3	07-499 (Upstate)	7	8.2	

The comparison of S5P and PRC associated markers at PRC-repressed genes, revealed that the occupancy of H3K27me3 and H2Aub1, and their respective enzymatic activities, Ezh2 and Ring1B, consistently mirrors that of S5P. At some genes the pattern of enrichment is more localized at the promoter region (*Gata4*), while at others high levels of S5P and PRC associated markers are detected throughout coding regions (*Atoh1, Msx1, Hoxa7*). The striking coincidence between PRC and RNAPII occupancy suggests that RNAPII and PRCs may simultaneously associate with the same chromatin, with PRCs traveling through genes. This topic will be subject of further investigation in Chapter 5.

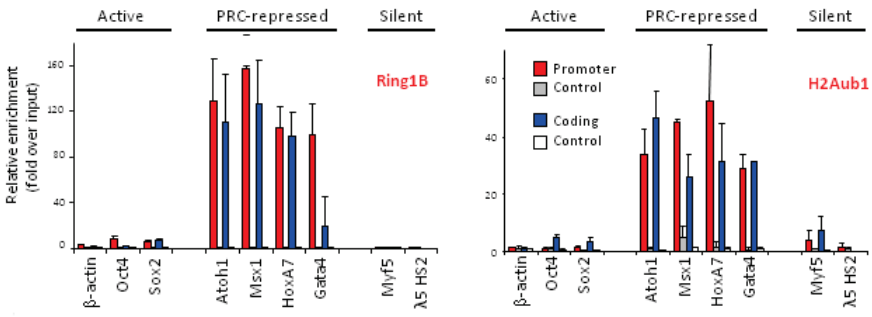


Figure 4.1: Quality control qPCR for H2Aub1 and Ring1B ChIP samples across a panel of control genes also depicted in Figure 4.2. Enrichment of Ring1B and H2Aub1 at active, PRC-repressed and silent genes in murine OS25 ES cells was assessed by ChIP and qPCR at promoter (red bars) and coding regions (blue bars), using published primers (Stock et al., 2007; qPCR experiments were performed by E. Brookes). ChIP enrichment is expressed relative to input DNA using the same amount of DNA in the PCR. Background levels (mean enrichment from control antibody and beads alone) at promoter and coding regions are shown as pale grey or white bars, respectively. Mean and standard deviations are presented from 2-4 independent ChIP experiments. The same ChIP samples were mapped by high-throughput sequencing.

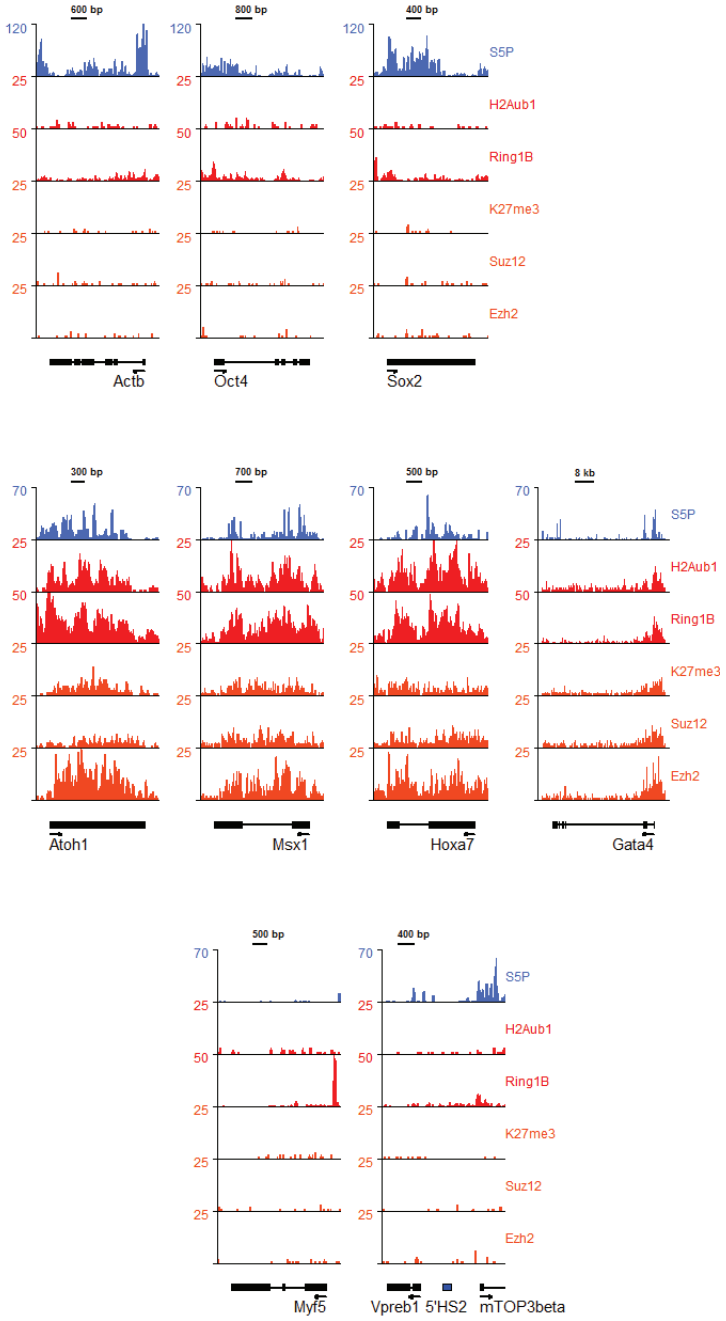


Figure 4.2: ChIP-Seq and mRNA-Seq profiles across single genes also depicted in depicted in Figure 4.1. S5P, H2Aub1 and Ring1B were sequenced in this study; H3K27me3, Suz12 and Ezh2 were retrieved from published studies (Mikkelsen et al. 2007, Ku et al. 2008). ChIP-Seq profiles are comparable to qPCR, with high levels of all PRC associated markers at PRC-repressed genes. Y-axes represent number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

4.2.2 *Chromatin and RNAPII patterns at Active, Bivalent and Inactive genes*

One of the main goals of our ChIP-Seq data analysis was to characterize different classes of genes according to their RNAPII, Polycomb and expression status. To achieve this aim it was essential to analyse datasets with adequate sequencing depth so that both the background and the binding sites are classified properly.

Genome-wide Ring1B occupancy data was previously published by Ku et al. [2008] (using the same antibody clone, A-20). Comparison of ChIP-Seq average profiles at highly ("Top20%") and least ("Bottom20%") expressed genes and at bivalent genes shows that, as expected, bivalent genes are enriched for PRC1 and PRC2 marks when compared to active or silent genes (Figure 4.3). Moreover, average ChIP-Seq profiles show that our Ring1B dataset has greater signal intensity than the previous dataset (compare the different y-axis scales in Figure 4.3). In addition, a sharp peak of Ring1B can be identified immediately upstream of the TSS of active genes (Figure 4.3; green lines) which is not seen in the published dataset with a lower depth of sequencing. Mapping of Ring1B in ES-ERT2 (conditional knockout of Ring1B) suggests that this peak may be unspecific (not shown).

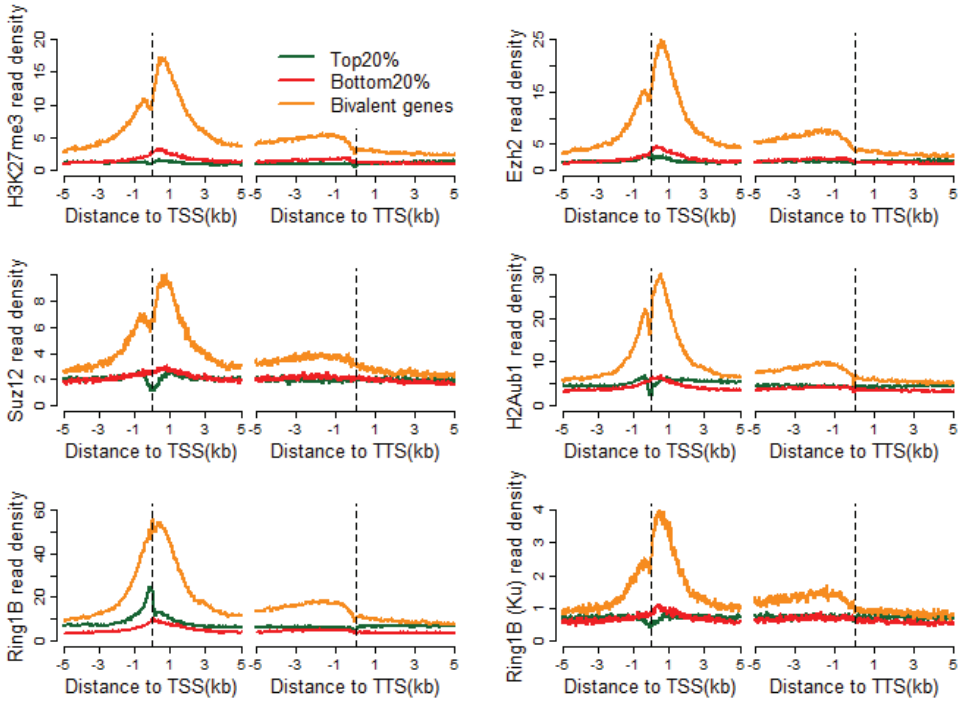


Figure 4.3: Genome-wide profiles of PRC related marks at most and least expressed and bivalent genes. Profiles of ChIP-Seq read density (depth per 10bp) profiles were computed for the region surrounding the TSS and TTS of highly (Top20%) and least (Bottom20%) expressed genes and bivalent genes. Top20% and Bottom20% were classified according to FPKM levels (section 3.3.5). Bivalent genes were classified in Mikkelsen et al. [2007]. The dashed line indicates the representative position of the TSS and TTS.

To compare the ChIP-Seq signal globally at gene promoters, we calculated the sequencing depth within ±1kb region centered at the TSS of 18860 genes and performed pairwise correlations (Figure 4.4). All 5 markers (H2Aub1, Ring1B, H3K27me3, Suz12, Ezh2) show positive correlation in the range from 0.38 to 0.68 (Spearman's rank correlation coefficient), confirming that, in the vast majority of cases, PRC2 and PRC1 proteins bind to a common set of promoters. Lower correlations are seen between the two Ring1B datasets (0.38) and between all correlations involving Suz12 (Mikkelsen et al., 2007) and published Ring1B (Ku et al., 2008) sets, possibly due to lower depth of signal in these published data sets. As expected, higher correlation coefficients (depicted in red; Figure 4.4) are seen between the catalytic PRC subunits (Ezh2, Ring1B) and the mark they catalyse (H3K27me3, H2Aub1, respectively; Spearman's correlation = 0.68). Interestingly, H2Aub1 and

H3K27me3 signals are also well correlated, contrasting to what is observed between Ring1B and H3K27me3.

As the published Ring1B dataset (Ku et al., 2008) was lower depth and our own dataset had a non-specific component, the downstream analyses relied on the H2Aub1 dataset, to most accurately describe PRC1 presence and activity. Moreover, relying on H2Aub1 as a measure of "PRC1 activity", is consistent with the analysis of PRC2 activity based on H3K27me3, as been done for many genome-wide studies of Polycomb function (Mikkelsen et al., 2007, Zhao et al., 2007, Wei et al., 2009, Pan et al., 2007).

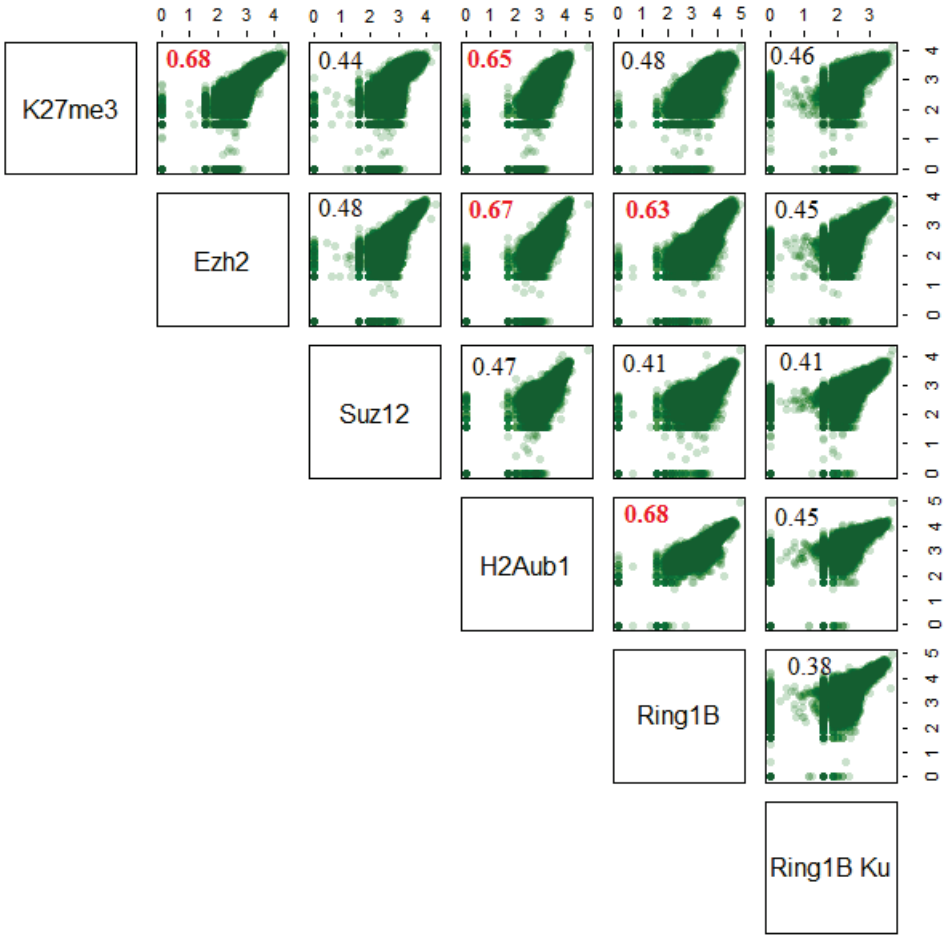


Figure 4.4: Pairwise correlation between levels of PRC markers, H2Aub1, Ring1B, H3K27me3, Suz12 and Ezh2 at promoters. Sequencing depth was calculated in 2kb windows centered at genes' TSSs. Scatter plots are displayed in the log10 scale for visualisation purposes. To avoid taking the log of 0, the value 1 was added prior taking the log. Speaman's correlation coefficients are shown in each plot.

4.2.3 *Classification RNAPII or PRCs ChIP-Seq signal*

To investigate the association of Polycomb-target genes with RNAPII complexes bearing specific CTD modifications, we used a methodology developed in collaboration with Sarah Teichman's group, MRC-LMB Cambridge. This methodology relies on the scoring of windows of interest (which are defined with respect to gene regions) based on read density and on the identification of a threshold which is used to classify the windows into "negative" and "positive", thus yielding estimates of the fraction of modified/bound genes above background. This methodology classifies each gene according to the presence or absence of RNAPII (S5P, 8WG16, S7P and S2P), H3K4me3, H3K27me3 and H2Aub1. This approach was recently published as a webtool (EpiCHIP; Hebenstreit et al., 2011).

The method involves four main processing steps; an overview of the analysis strategy is given in Figure 4.5. (i) Firstly, genic regions of interest (5′ or 3′ end of genes) were identified. To do so, I aligned RefSeq genes according to their annotated TSS or TTS and produced genome-wide average read distributions. Genic regions of interest were chosen accordingly to where RNAPII and Polycomb was present: a 2kb window centered at the TSS was chosen for S5P, 8WG16, S7P, H3K4me3, H3K27me3 and H2Aub1, a 2kb window downstream the TTS for S2P, and finally a 2kb window upstream of the TTS was also considered for S5P data analysis. (ii) Secondly, once windows of interest were identified, the sequencing depth (i.e. number of read nucleotides that fall into that window) was calculated to score every window. At this step, because all windows have the same length (2kb), it was not necessary to normalise depth counts to window width. (iii) Thirdly, density plots were used to display the global distribution of ChIP-Seq sequencing depth. When displayed on a log scale, the bimodality of the data was typically emphasized, suggesting that it is composed of two subpopulations, each with separate, but overlapping distributions. (iv) Fourthly, the two subpopulations were classified into "noise" (background) and "signal", using an expectation maximization algorithm (optim function in R; Nelder and Mead, 1965, R Development Core Team, 2010) to fit a two-component mixture model of combinations of Normal and lognormal distributions. This approach allows us to identify a threshold at a specific false discovery rate (FDR) which is then used

to categorize window depth values into binary "positive" or "negative" (1,0) events that denote, respectively, the presence or absence of a given factor (either at TSS or TTS region). We chose an FDR of 1% to perform the binary classifications detailed below. Finally, windows with scores above the identified threshold were classified as "signal", and thus the corresponding gene was classified as being positively associated to a given marker at the genomic window of choice; while windows with depth scores that fall below the identified threshold were classified as "background". This analysis produced a table of (1,0) scores for all RefSeq genes and markers (S5P, 8WG16, S7P, H3K4me3, H3K27me3 and H2Aub1).

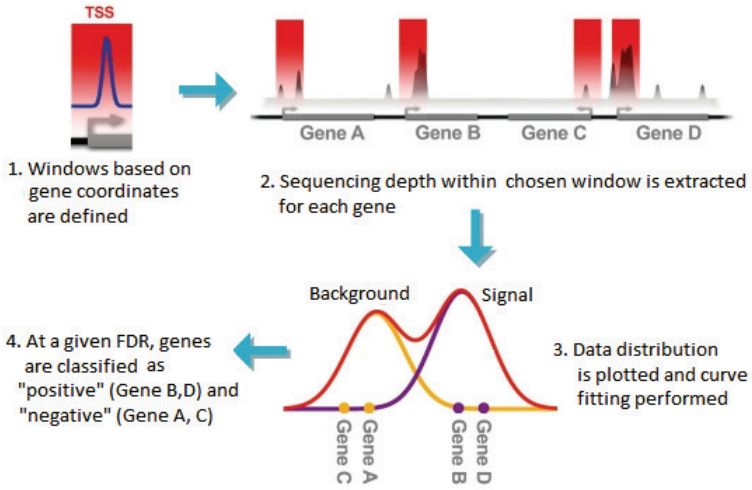


Figure 4.5: Overview of the classification strategy. The image was adapted from Hebenstreit et al., 2011

4.2.3.1 *(i) Genome-wide average profiles*

To identify optimal windows of marker enrichment, the average ChIP-Seq profiles across -5kb to +5kb regions relative to the TSS and TTS were calculated. Average profiles were calculated by aligning all RefSeq genes relative to their TSS and TTS and by measuring the sequencing depth at sliding windows of 10bp width. Figure 4.6 shows average profiles for all RefSeq genes.

4.2.3.2 *(ii) Identification of the optimal windows for classification*

Careful analysis of the average read profiles at TTS and TTS indicate indicate that the most prominent peaks are detected at the promoter region for S5P, 8WG16, S7P, H3K4me3, H3K27me3 and H2Aub1 datasets and downstream of the TTS region for the S2P dataset. The approximate size of the peak width is 2kb, and therefore, we have chosen the 2kb windows depicted in Figure 4.6 for downstream analysis. For the S5P dataset, a secondary peak was detected for S5P upstream the 3′ end of genes which we are interested in measuring to identify genes in which S5P is observed throughout the gene's coding region, therefore the 2kb region upstream of the TTS was also classified. In contrast, our ChIP-Seq datasets for the mock IP control distribution is much flatter, showing only a slight increase at the TSS.

The size of the windows was chosen in a heuristic way to ensure a clear separation of "real signal" from "noise". The precise size and position of the chosen window is not crucial for the downstream analysis, as long as it represents genomic sites where the separation between signal and background is clear; for instance we have already observed that inactive genes (such as those previously classified as bottom 20%) are clearly depleted of S5P at their promoters and S2P at their 3′end, while at active genes S5P and S2P are clearly enriched (Figure 3.7; Hebenstreit et al., 2011). Nonetheless, to investigate the effect of window size the classification strategy was repeated using a range of window sizes (described below in subsection 4.2.4).

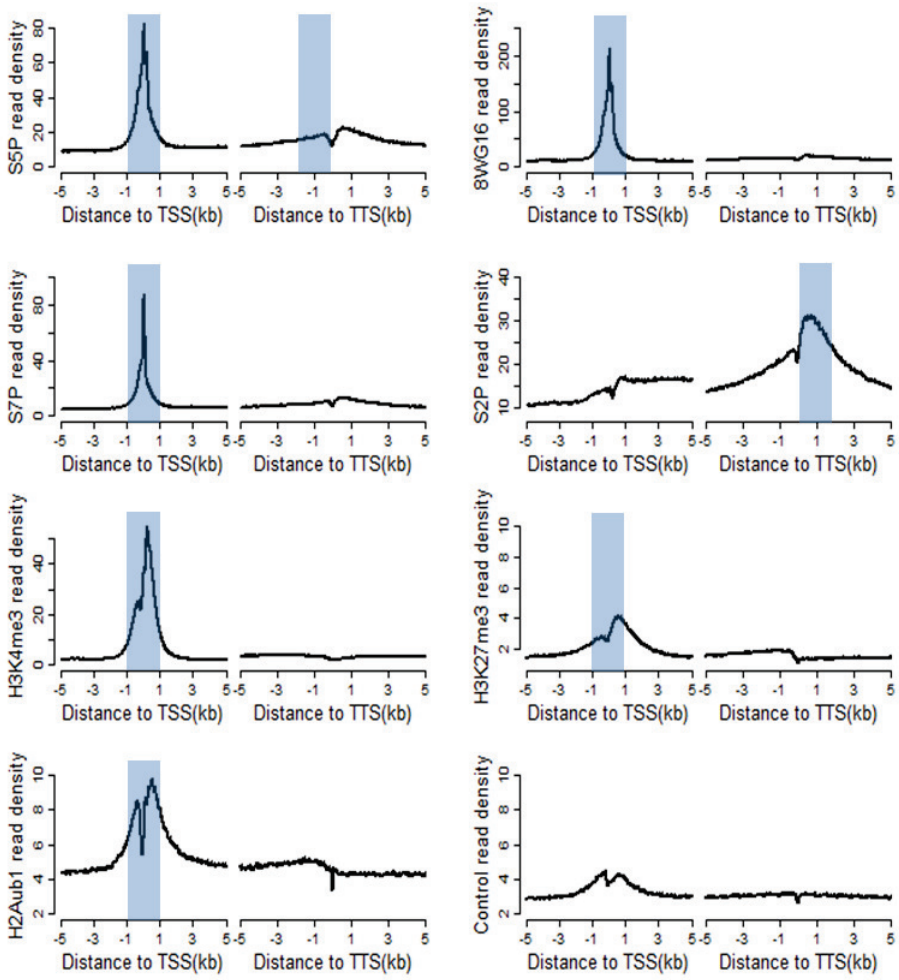


Figure 4.6: Genome-wide average read profiles at RefSeq genes. Average ChIP-Seq read density (depth per 10bp) profiles at RefSeq genes are shown for RNAPII (S5P, 8WG16, S7P and S2P) and H3K4me3, H3K27me3 and H2Aub1 datasets to display the general binding pattern. S5P, 8WG16, S7P, H3K4me3, H3K27me3 and H2Aub1 show a strong enrichment at 2kb region centered at the TSS, while S2P shows a stronger enrichment at 2kb region downstream of the TTS. A secondary 2kb peak upstream of the TTS is also considered for downstream S5P data analysis. The control is a mock IP experiment with no antibody. Shaded windows represent the chosen 2kb windows for the classification strategy.

4.2.3.3 *(iii) Bimodality of the distribution of Chip-Seq signal*

To determine whether the distribution of ChIP-Seq signal at windows consist of a mixture of two subpopulations (noise and signal), the global distribution was displayed as a density plot for all analysed variables (Figure 4.7). In every case, the distribution of values is skewed with a long tail to the right. The skewness of the distribution is less accentuated for the H3K27me3 and H2Aub1 datasets which

means that relatively fewer genes are represented by high window scores. This is in agreement with what is expected according to the overall reduced spread of H3K27me3 and H2Aub1 marks along the entire genome when compared to RNAPII binding or the H3K4me3 mark. The longer tail is also observed for the control data (Figure 4.8), but values are relatively more evenly distributed on both sides of the mean then for other ChIP-Seq datasets. The bimodality of the distribution is emphasized by log-transforming the data which demonstrates, in some cases very clearly, that the sample data can be modeled by the sum of two separate but overlapping components, one corresponding to low values ("noise") which is also well modeled by the control sample, and another one corresponding to higher values ("signal").

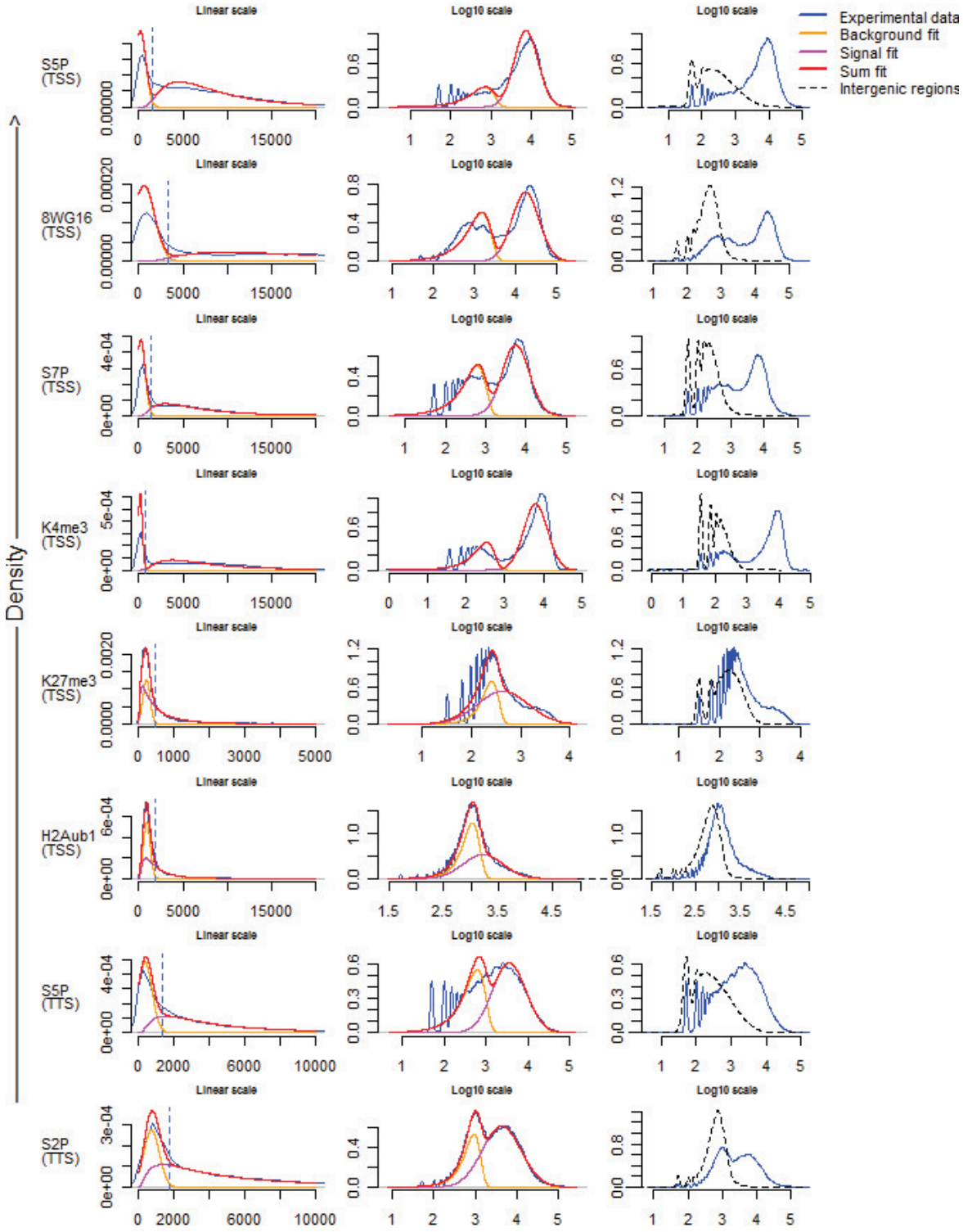


Figure 4.7: Curve fitting of RNAPII (S5P, 8WG16, S7P, S2P), H3K4me3, H3K27me3 and H2Aub1 data. Kernel density estimates of the distributions of sequencing depths at 2kb windows (for all RefSeq genes) are shown in the linear (left) and log10 (middle, right) scales (blue lines). A combination of a Normal and a lognormal distribution fitted to each dataset are overlaid (red, orange and pink lines). The blue dashed lines on the left plots indicate the thresholds at FDR=0.01. On the right, dotted black lines represent the signal distribution of random intergenic regions of 2kb size. Parameters and alternative fits are shown in Table 4.2.

A. Modeling the noise

To investigate whether the left peak seen in the log representation of the data corresponds to background signal, 18860 intergenic windows of 2kb each were randomly selected and plotted. For the mock-IP control, the distribution of intergenic window scores largely overlaps with the distribution at TSS (not shown), whereas for all the other variables the resulting distribution overlaps the left peak suggesting that this left peak largely consists of experimental and biological noise (Figure 4.7, dotted lines).

The shape of the distribution for the mock IP sample suggests that a Normal distribution would fit the distribution. To compare the fit of a Normal curve with other possible alternatives (lognormal or truncated Normal), we used Bayesian information criterion (BIC) as our goodness-of-fit test. The results indicate that Normal and truncated Normal distributions fit the data better then lognormal (Figure 4.8).

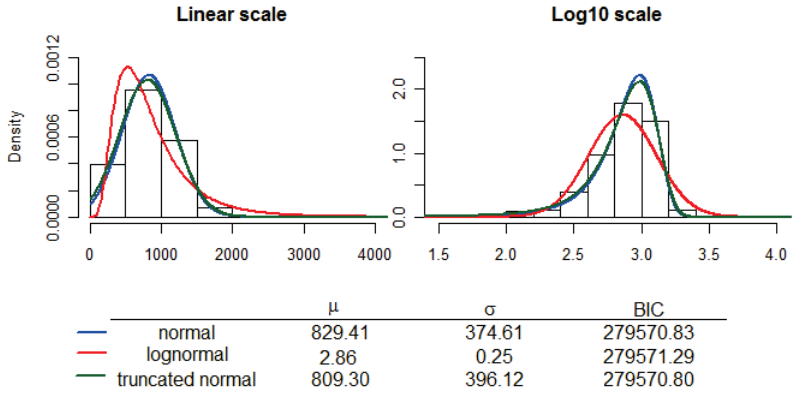


Figure 4.8: Curve fitting of the mock IP control distribution. The distribution of sequencing depth calculated at 2kb windows centered at the TSS for all RefSeq genes is shown as an histogram in the linear (left) and log10 (right) scales. Maximum likelihood fits of Normal (blue), lognormal (red) and truncated Normal (green) are overlaid. Parameters of the fitted distributions and BICs are also represented.

B. Modeling the two-component distribution

To fit a two-component mixture distribution, we used an expectation maximization algorithm (optim function in R; Nelder and Mead, 1965, R Development Core Team, 2010) and performed several fittings of the data using various combinations of

Normal and lognormal distributions (lognormal - lognormal; Normal - lognormal; truncated Normal - lognormal).

Modeling the data with two lognormals led to the estimation of very high thresholds, mainly due to the fact that the "noise" distribution is poorly represented by a lognormal curve (see for example H3K4me3 and S5P fittings Figure 4.9). In this case, the heavy right tail of the lognormal curve leads to an increased estimation of false positives (FDR = 1%), and thus to an increased unrealistic threshold that separates "noise" from "signal".

Both of the other two tested models (truncated Normal - lognormal; Normal - lognormal) fit the data well, but the best representation of the shape of the distribution was achieved for the two-component mixture composed of a truncated Normal distribution and a lognormal distribution. In the majority of the cases, the truncated Normal - lognormal also gave best BIC, although the differences were very moderate between the three models. Best-fit results are in agreement with what has been seen for H3K9/14ac and H3K4me3 data, where the fit of background and signal models of windowed data were studied extensively (Hebenstreit et al., 2011). All fittings and parameters are presented in Table 4.2.

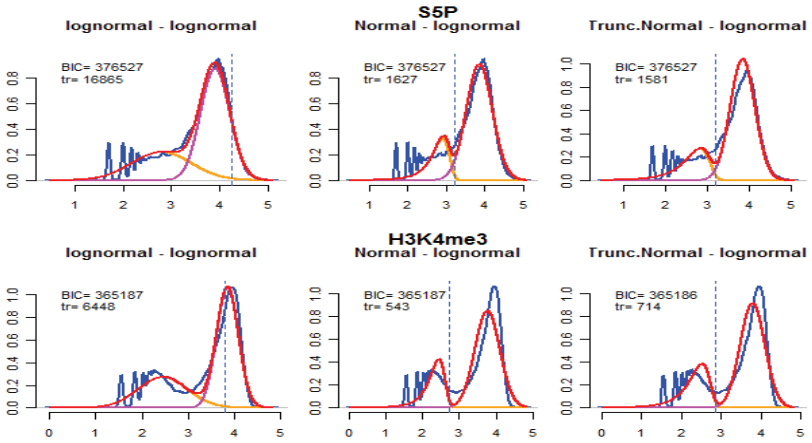


Figure 4.9: Example fitting of 3 models to H3K4me3 and S5P data sets. Kernel density estimates of the distributions of sequencing depths at 2kb windows (for all RefSeq genes) are shown in the log10 scale. A combination of a Normal and a lognormal distribution fitted to each dataset are overlaid (red, orange and pink lines). Blue dashed lines represent the estimated threshold ("tr"). The lognormal-lognormal fittings leads to an overestimation of the threshold. Both the other two models (Normal - lognormal; truncated Normal - lognormal) fit the data well. Parameters and alternative fits are shown in Table 4.2.

Table 4.2: Parameters of the fitted distributions for different window sizes. Several fittings of the data were performed using various combinations of Normal and lognormal distributions (truncated Normal - lognormal; lognormal - lognormal; Normal - lognormal). The best representation of the shape of the experimental distribution was achieved for the two-component mixture composed of a truncated Normal and a lognormal distribution and is represented in Figure 4.7.

	Sample	window	α	μ_0	σ_0	μ_1	σ_1	threshold	BIC
lognormal + lognormal	S5P	2kb	0.3	2.8	0.6	3.9	0.3	16865.1	376527
	8WG16	2kb	0.5	3.1	0.5	4.3	0.3	29691.0	416055
	S7P	2kb	0.5	2.7	0.5	3.8	0.3	7099.1	374074
	K4me3	2kb	0.4	2.5	0.5	3.9	0.2	6447.7	365187
	K27me3	2kb	0.3	2.4	0.2	2.6	0.5	814.6	306235
	H2Aub1	2kb	0.5	3.0	0.2	3.1	0.4	3120.7	333362
	S5P (TTS)	2kb	0.3	2.4	0.4	3.4	0.5	3062.0	361981
	S2P	2kb	0.1	2.9	0.2	3.5	0.6	2024.9	398565
normal + lognormal	S5P	2kb	0.2	568.6	452.8	3.9	0.3	1626.9	376527
	8WG16	2kb	0.4	1016.3	681.8	4.2	0.4	2607.2	416056
	S7P	2kb	0.4	599.0	381.8	3.7	0.3	1492.6	374072
	K4me3	2kb	0.3	207.2	144.3	3.8	0.3	543.1	365187
	K27me3	2kb	0.3	223.5	85.1	2.6	0.5	444.6	306232
	H2Aub1	2kb	0.5	958.8	359.0	3.2	0.4	1918.3	333362
	S5P (TTS)	2kb	0.3	362.0	241.9	3.5	0.4	938.6	361963
	S2P	2kb	0.2	753.4	332.5	3.5	0.5	1564.8	398564
truncated normal + lognormal	S5P	2kb	0.2	217.1	585.0	3.9	0.3	1581.0	376527
	8WG16	2kb	0.4	608.7	1180.6	4.2	0.3	3363.8	416056
	S7P	2kb	0.4	250.4	464.5	3.7	0.3	1337.8	374074
	K4me3	2kb	0.3	183.3	228.3	3.8	0.3	714.4	365186
	K27me3	2kb	0.3	219.5	100.1	2.6	0.5	479.6	306232
	H2Aub1	2kb	0.5	916.1	382.5	3.2	0.4	1922.9	333362
	S5P (TTS)	2kb	0.4	352.8	402.9	3.5	0.4	1313.5	361966
	S2P	2kb	0.3	722.8	434.9	3.6	0.5	1771.7	398563
	S5P	1kb	0.2	165.3	308.4	3.7	0.4	886.2	352315
	8WG16	1kb	0.4	212.4	789.5	4.1	0.3	2053.9	402819
	S7P	1kb	0.4	3.6	403.5	3.6	0.3	946.3	349031
	K4me3	1kb	0.3	49.9	192.0	3.6	0.3	497.0	339816
	K27me3	1kb	0.3	100.2	57.1	2.3	0.5	248.9	269132
	H2Aub1	1kb	0.4	402.7	222.1	2.8	0.4	1006.6	309008
	S5P (TTS)	1kb	0.3	181.0	145.8	3.3	0.4	525.6	321791
	S2P	1kb	0.2	605.3	220.3	3.2	0.6	1163.0	377827
	S5P	500bp	0.3	-679.5	642.9	3.6	0.3	824.5	322278
	8WG16	500bp	0.4	226.4	402.9	3.9	0.4	1166.4	381893
	S7P	500bp	0.4	-27.2	307.8	3.5	0.3	692.9	317781
	K4me3	500bp	0.3	50.8	118.7	3.3	0.3	327.3	301241
	K27me3	500bp	0.4	63.8	41.2	2.1	0.5	172.9	213928
	H2Aub1	500bp	0.8	176.3	161.6	2.9	0.2	566.5	270424
	S5P (TTS)	500bp	0.6	242.5	248.5	3.2	0.3	842.8	264927
	S2P	500bp	0.8	425.1	596.6	3.6	0.4	1859.0	343709

4.2.3.4 *(iv) Threshold calculation*

One advantage of the curve fitting approach is the derivation of a threshold at given false discovery rate (FDR) whilst retaining a quantitative raw measure of read density at the analysed windows. The FDR is the quotient between the expected proportion of false positive detections and the total number of detections. We used an FDR of 1%, thus setting the threshold to a value where the proportion of 'noise' corresponds to 1% of the total area of the mixture on the right of the threshold, ensuring that the number of false positive detections will be small. The threshold was then used to categorise the continuous values into binary 'positive' or 'negative' (1,0) events that denote, respectively, the presence or absence of a given histone or RNAPII mark (either at TSS or TTS region).

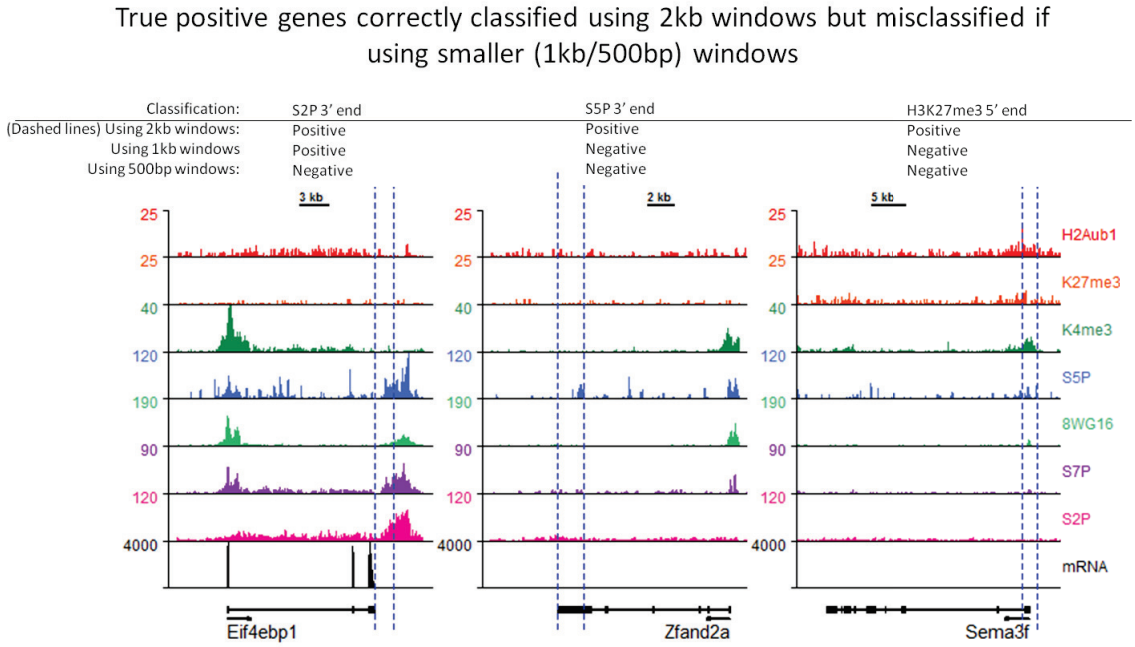
4.2.4 *Robustness of the classification*

To investigate the robustness of the method against different window settings, the classifications were performed using different window sizes (the parameters identified for each fitting are represented in Table 4.2). Windows of 1kb (using the truncated Normal/lognormal model at FDR = 1%) give similar results to windows of 2kb. On average 94% of genes are commonly classified as positive between 1kb and 2kb window approaches. Using windows of 500bp also leads to similar results to windows of 2kb, where 64% to 97% (86% on average) of classifications agree between the two approaches (see Table 4.3 for the number of genes which agree between classifications). The largest differences in the total "positive" / "negative" calls are seen when using windows of 500bp at the TTS to classify genes according to their S5P and S2P signal (Table 4.3).

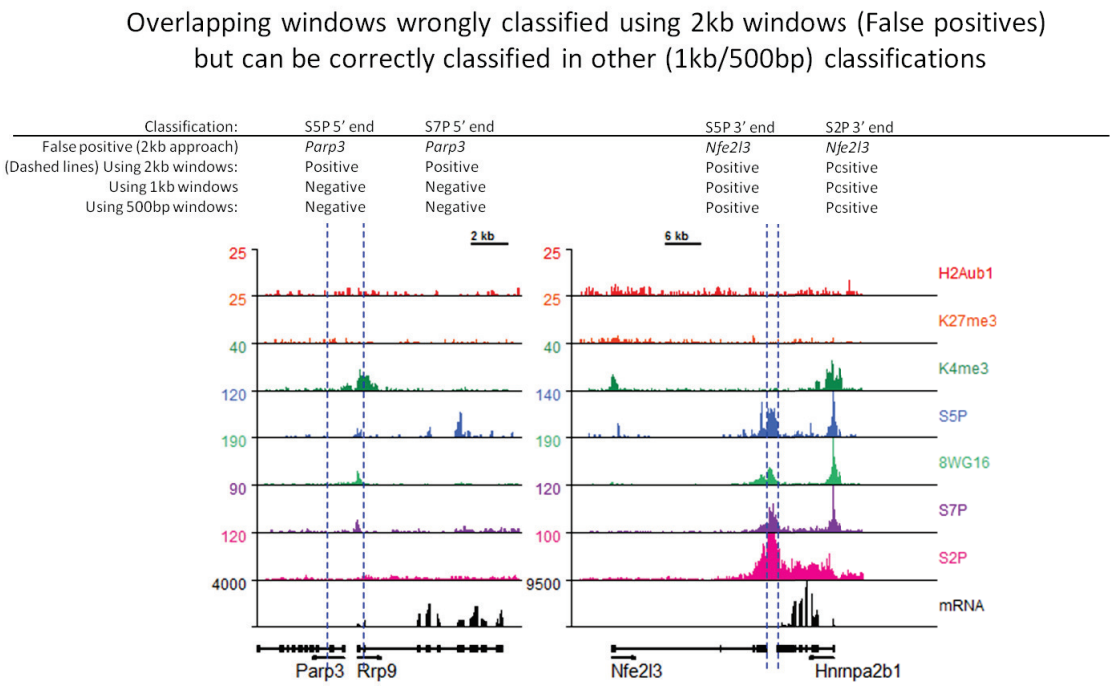
Table 4.3: Agreement between classifications using different windows sizes. The number of "positive" and "negative" calls commonly identified using either 1kb or 500bps window sizes compared to the ones obtained using 2kb windows are shown. The percentage of common classifications represents the total number of positive and negative calls out of the total number of windows (n=18860).

	Nr of "positive" windows			Nr common "positive" classifications		Nr common "negative" classifications		% of common classifications	
Sample	2kb	1kb	500bp	1kb	500bp	1kb	500bp	1kb	500bp
S5P (TSS)	13457	13364	11774	12966	11615	5005	5244	95%	89%
8WG16	11093	11156	11200	10904	10789	7515	7356	98%	96%
S7P	10795	10532	10141	10412	10018	7945	7942	97%	95%
K4me3	13188	13005	12661	12939	12599	5606	5610	98%	97%
K27me3	5890	5767	4047	5093	3767	12296	12690	92%	87%
H2Aub1	3840	3839	2936	3258	2452	14439	14536	94%	90%
S5P (TTS)	9650	9626	4583	8453	4507	8037	9134	87%	72%
S2P (TTS)	10601	10133	3871	9309	3871	8142	8259	93%	64%

Close inspection of genes differentially classified between the three methods (using 2kb, 1kb or 500bp windows) revealed cases in which low depth signal is not detected if one uses small window sizes (500bp) as opposed to bigger windows (1kb or 2kb). Concrete examples are given in Figure 4.10a for *Eif4ebp* and *Zfand2a* genes which are wrongly classified as S2P-negative or S5P-negative, respectively, if one uses 500bp windows, but correctly classified as "positive" if one uses a bigger window size (2kb) defined relatively to the 3′ end. Similar examples of broad patterns which are wrongly classified using smaller (500bp) windows include those of the H3K27me3 mark at the promoter of *Sema3f* gene (Figure 4.10a). Nonetheless, for the great majority of the cases, our results show that changing the window sizes results only in slight differences in the classification of genes.



(a)



(b)

Figure 4.10: Example of genes differentially classified between three approaches using three different window sizes (2kb, 1kb and 500bp). (a) Broad patterns at *Eif4ebp*, *Zfand2a* and *Sema3f* are positively detected using 2kb window sizes, but lost if using smaller windows (500bp). (b) Overlapping windows lead to the misclassification of genes. *Parp3* and *Rrp9* are closely located, but only *Rrp9* carries S5P enrichment at its promoter. *Nfe2l3* is misclassified independently of the window size chosen due to overlapping signal from *Hnmpa2b1*.

Other examples of differentially classified genes includes cases in which the use of smaller windows provides better resolution, two examples are depicted in Figure 4.10b. For example, the 2kb windows centered at the TSS of *Parp3* and *Rrp9* overlap and thus *Parp3* becomes wrongly classified as "S5P-positive" when in fact the signal clearly belongs to *Rrp9* promoter window. However, when using 1kb or 500bp promoter windows the signal is separated, and *Parp3* is correctly classified as "S5P-negative" (Figure 4.10b). The same type of example is shown for *Nlf2l3* gene (Figure 4.10b), which is wrongly classified by all approaches (2kb, 1kb or 500bp windows) due to signal from the nearby *Hnmpa2b1* gene.

For downstream analysis, classifications obtained for 2kb windows were used. Since several cases were identified where two windows belonging to different genes overlap, we removed genes where windows positive for a given marker overlap by more than 10% (i.e. 200 bp). The number of genes that were removed in each dataset is represented in Table 4.4.

Table 4.4: Number of overlapped windows excluded from each data set. If two or more "positive" windows from distinct RefSeq genes overlap it is not possible to declare to which gene (or if to both) the observed signal belongs. To avoid misinterpretation of the results, "positive" windows that overlapped by at least 10% (i.e. 200 nucleotides) were ignored. The resulting number of ignored classifications is represented.

	S5P (TSS)	8WG16	S7P	K4 me3	K27 me3	H2Aub1	S5P (TTS)	S2P
Nr of windows removed from the classification	1802	1766	1760	1814	218	172	1785	1949

4.2.5 *Validation of the classification by comparing with previous classifications*

We compared this novel classification strategy with other published classifications, namely for 'bivalent' and 'PRC1-positive bivalent' genes (Ku et al., 2008, Mikkelsen et al., 2007, respectively). We identify 4115 genes bound by H3K4me3 and H3K27me3 and, from these, the majority is also H2Aub1+ (H3K27me3+ H3K4me3+ H2Aub1+; n=2863). These novel analyses identify larger numbers of 'bivalent' (n=4115) and 'PRC1-positive bivalent' (n=2863) genes than previous studies (Figure 4.11).

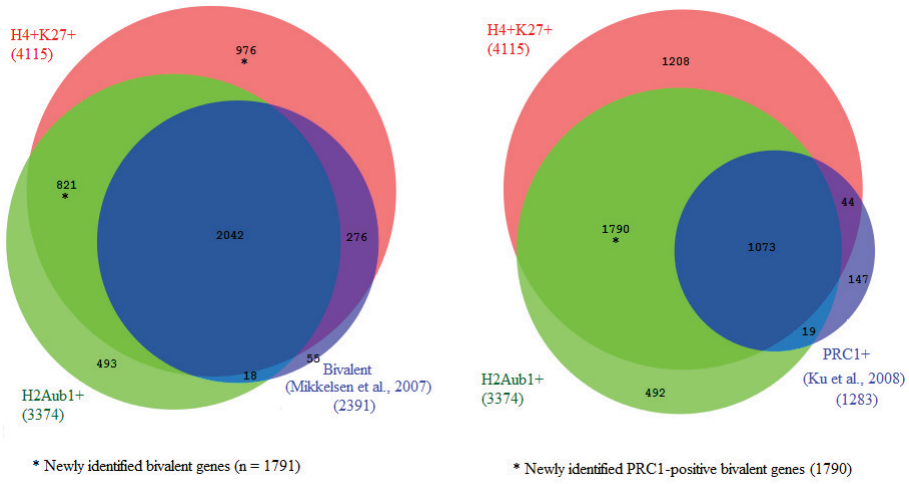
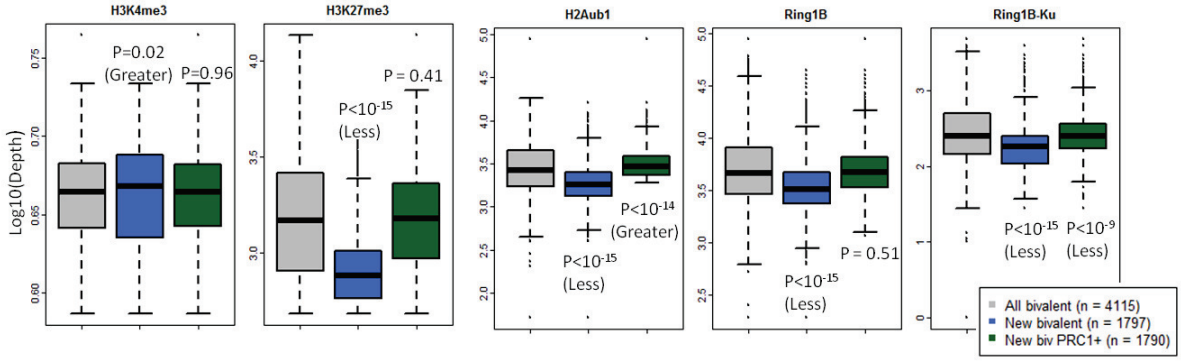


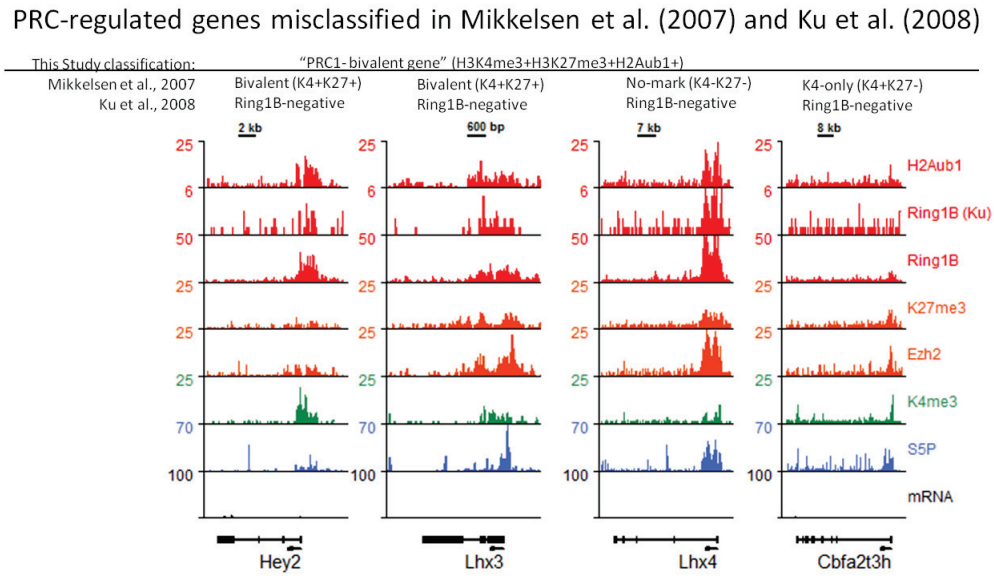
Figure 4.11: Bivalent gene classifications in the current and previous studies. We identify 4115 bivalent genes (K4+K27+), of which 44% (n = 1797) were not previously classified as bivalent domains (Mikkelsen et al., 2007). 46% of newly identified bivalent genes (821 out of 1797) are also H2Aub1+. Compared to Ku et al., 2008 classifications, we identify many newly 'PRC1-positive bivalent' genes (n=1839), probably due to classification strategy and the lower depth of published Ring1B data compared to our new H2Aub1 data.

The differences in classifications arises mainly from (i) more genes previously classified as 'K4-only' (H3K4me3+H3K27me3-) now being classified as marked with H3K27me3 and/or H2Aub1; (ii) many genes, previously classified as PRC1-negative now being classified as H2Aub1+ genes. Inspection of ChIP-Seq levels confirms that newly classified bivalent genes have on average lower H3K27me3 signal (p-value < 10^{-15}; one-tailed Wilcoxon rank-sum test), and thus could have been missed from Mikkelsen et al. [2007] classifications. Using the Ring1B dataset from Ku et al. [2008], we verify that the newly classified "PRC1-positive bivalent" have lower Ring1B signal when compared to all bivalent genes (p-value < 10^{-9}; one-tailed Wilcoxon rank-sum test); but when using our own, higher depth, Ring1B and H2Aub1 datasets the same result is not confirmed (Figure 4.12a). Indeed, the subclass of newly identified PRC1-positive bivalent genes has significantly higher levels of H2Aub1 when compared to all bivalent genes (p-value < 10^{-15}; one-tailed Wilcoxon rank-sum test). In contrast to the previous classification of PRC1-bound bivalent genes (based on Ring1B ChIP; Ku et al., 2008), we identify that the majority of bivalent genes are marked by H2Aub1. For instance, according to our own datasets, genes such as those represented in Figure 4.12b are classified as bivalent Ring1B-negative (*Hey2*, *Lhx3*) in

Ku et al. [2008] but have detectable Ring1B and H2Aub1 signal. Other genes (*Lhx4*, *Cbfa2t3h*) are classified as H3K27me3 negative by Mikkelsen et al. [2007] and Ring1B-negative by Ku et al. [2008], but low depth signal is seen when using the data sets from these same studies (Figure 4.12b). Differences between ours, Mikkelsen et al. [2007] and Ku et al. [2008] classifications could arise due to greater longevity of the nucleosome modification (H2Aub1) than Ring1B at these promoters, due to the higher signal produced by ChIP with histone antibodies than with Ring1B, or due to difficulties in detecting lower Ring1B or H3K27me3 signals. Nevertheless, our evidence suggests that the majority of bivalent genes are also PRC1 targets (70%; based on H2Aub1 annotation; 2863 out of 4115 bivalent genes). In addition, almost all H2Aub1 marked genes (n = 3374) are marked by H3K4me3 (n = 3287; 97%), while the same is not true for the H3K27me3 mark, which can more commonly be found in the absence of H3K4me3 (79% of H3K27me3 are also H3K4me3 marked). This interesting finding indicates that the H2Aub1 mark more faithfully represents bivalent chromatin then H3K27me3 mark.



(a)



(b)

Figure 4.12: Comparison of H3K4me3, H3K27me3 and H2Aub1 sequencing depth at the promoters of genes that differ between the current and previous classifications. (a) Newly classified bivalent genes have on average lower H3K27me3 and PRC1 (H2Aub1, Ring1B) signals, when compared to all bivalent genes. Newly identified "PRC1-positive bivalent" genes have lower Ring1B signal according to Ku et al. [2008] data set, but not according to our own data sets (H2Aub1 or Ring1B), when compared to all bivalent genes. Sequencing depth was calculated within ±1kb region centered at the TSS of all represented genes. Wilcoxon rank-sum test p-values are represented under each box plot for comparisons against "All bivalent genes". (b) ChIP-Seq and mRNA-Seq profiles across three example genes which are misclassified as PRC1-negative according to Ku et al. [2008] due to low Rin1B depth (see "Ring1B (Ku)"). Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

4.2.6 *Characterization of PRC+ genes*

The classification strategy based on measuring signal at windows of interest allowed us to identify a large cohort of PRC^+ genes (n=5628) in ES cells, either associated with both H2Aub1 and H3K27me3 (n=2931), only H3K27me3 (n=2254), or only H2Aub1 (n=443).

The PRC^+ cohort contains not only silent genes, but also genes with intermediate and high expression. The large range of expression levels at PRC^+ genes is also seen using published mRNA datasets (Figure 4.13) determined with different methodologies and ES cell lines. Heatmaps representing RNAPII modification at PRC^- and PRC^+ genes ordered by expression show that expression levels at PRC^+ genes are positively associated with presence of 8WG16, S7P and S2P (Figure 4.14), meaning that expression of PRC^+ genes is consistent with the the degree of RNAPII enrichment and its configuration. Least expressed PRC-target genes were generally associated with S5P only, whereas actively-expressed PRC-targets were occupied by both S5P and S2P, demonstrating genome-wide association between PRCs and RNAPII-S5P.

Our data reveals a considerable proportion of PRC-repressed genes that are transcriptionally active, to the level of what appears to be efficient transcription elongation (as seen from the enrichments on S2P and H3K36me3), but it also raises important questions on the interpretation of the observed coincidence between PRC-repression and efficient gene transcription.

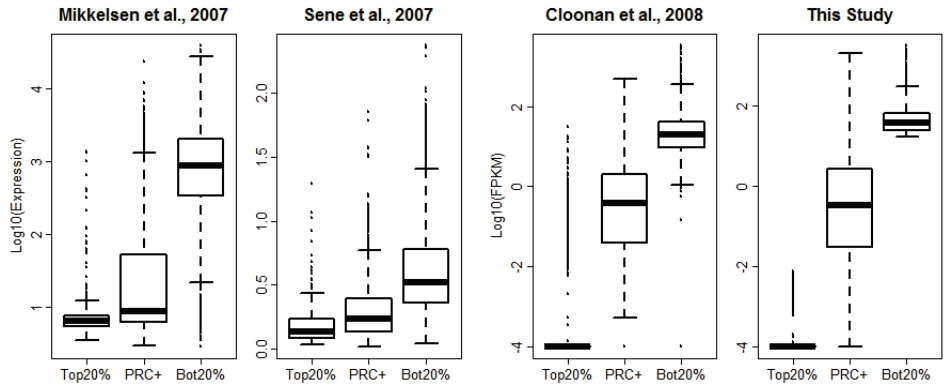


Figure 4.13: mRNA levels at PRC$^+$ genes encompass the spectrum of expression levels. mRNA expression levels were determined for the group of 20% highest expressed (Top20%), 20% lowest expressed (Bot20%) and PRC$^+$ genes using published mRNA-seq (this study, and Cloonan et al. [2008]) and microarray data (Mikkelsen et al., 2007, Sene et al., 2007). For visualisation purposes, expression levels are represented in log10 scale. To avoid taking the log of zero, a pseudo value of 0.0001 (in the case of FPKM) was added prior taking the log. Top20% and Bot20% were classified according to FPKM levels. PRC$^+$ genes show a wide range of expression levels.

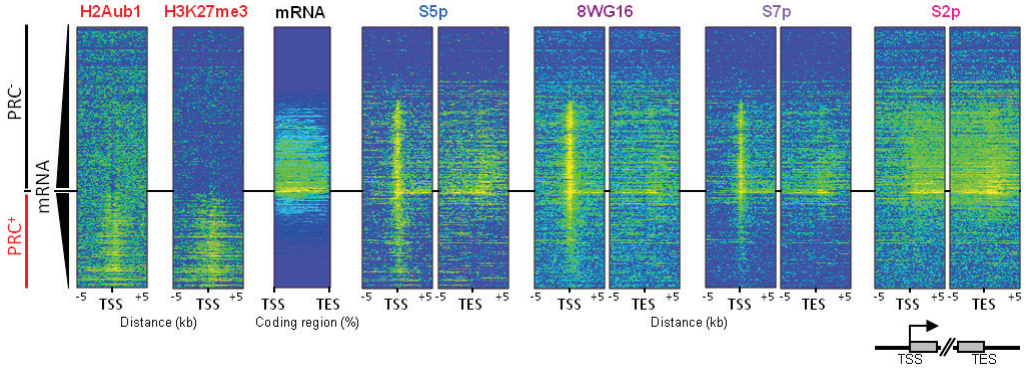


Figure 4.14: PRC-targets associate with different RNAPII modifications and expression levels. ChIP-Seq density heatmaps for 15404 non-overlapping RefSeq genes within 10kb windows centered on TSS or TTS (histone and RNAPII markers), and mRNA-Seq maps across coding regions normalized to length. PRC-negative (PRC-; n=9776) and PRC-positive (PRC+; n=5628) genes were classified and ordered by mRNA-seq levels. PRC+ genes include both silent and active genes, and associate with S5P irrespective of activity.

4.3 DISCUSSION

PRCs are important regulators in ES cells where they are known to silence developmental regulator and lineage-specific genes and to maintain the pluripotent state (Boyer et al., 2006, Lee et al., 2006).

We have generated ChIP-Seq genome-wide maps of Ring1B and the histone modification it catalyses, H2Aub1, and analysed publicly available ChIP-Seq datasets of Polycomb marks (Ezh2, Suz12 and Ring1B) and H3K27me3. Genome-wide analyses show that PRC-targets are strongly enriched for Polycomb proteins and the respective epigenetic marks they catalyse, H3K27me3 and H2Aub1, and that most PRCs co-associate with RNAPII at promoters and throughout coding regions (Figure 4.3). We have developed a strategy to classify genes according to the enrichment patterns of several RNAPII CTD phosphorylated residues, and the epigenetic marks H3K27me3 and H2Aub1. The classification scheme used here can be useful for the classification of other ChIP-Seq samples that have important roles in transcriptional control. Although the mechanisms and function of Polycomb complexes has been studied widely, the detailed relationship between RNAPII presence and Polycomb repression has never been documented in a genome-wide study.

4.3.1 *Distinguishing marked and bound genes from background*

A common feature of RNAPII and histone modification type of ChIP-Seq data is the fact that clear patterns of enrichment are distributed along target genes. The classification approach used here (based on a strategy developed in Hebenstreit et al., 2011) takes advantage of the fact that most of the genomic enrichment is expected to exist within consistent regions of genes. For instance H3K4me3 is primarily enriched at the TSS of bivalent and active genes, while S2P appears to mark more prominently the 3'-proximal region of transcribed genes (Figure 3.7, Figure 3.9). We demonstrate here a novel way for analyzing ChIP-seq data for RNAPII and histone modifications. This approach relies on measuring the ChIP-Seq signal in fixed windows for each gene, and in the distinction between background "noise" and biologically meaningful signal. In order to determine contributions of the background and real signal, a two-component mixture model is fitted to the distribution of the data by using the expectation maximization algorithm. The advantage of this approach is the derivation of a probabilistic criterion for identifying a threshold at a specific false discovery rate (FDR) allowing us to define which genes are positively associated to a given mark/protein and to quantify that signal at each individual gene.

4.3.2 *Improvements on the classification of bivalent and PRC1-positive genes*

Using the novel classification strategy, we identify a larger number of bivalent genes (H3K27me3+H3K4me3+; n=4115). Our classification recovers the majority of previously identified bivalent genes published in Mikkelsen et al. [2007](n = 2627; Figure 4.11). The newly classified bivalent genes have lower H3K27me3 and H2Aub1 signals (Figure 4.12a); these were probably missed from other classifications due to insensitivity to lower signals, in particular lower H3K27me3 signal. In a recent study (Young et al., 2011), the local pattern of enrichment of H3K27me3 in ES cells was extensively studied and genes were classified using specific methodology that took into account "broad" and "sharp-type" of H3K27me3 patterns of enrichment. The authors identified a total of 4495 H3K27me3 positive genes; of these, 3369 match our own H3K27me3 classification (65%).

Moreover, using the same approach, we identify more 'PRC1-positive bivalent' genes (H3K27me3+ H3K4me3+ H2Aub1+; n=2863) than previous studies (n = 1839; 64%). We have found that bivalent genes previously classified as PRC1-negative (Ku et al., 2008) exhibit lower levels of Ring1B signal based on the published dataset (Ku et al., 2008); indicating that they could have been missed from previous classification due to the insensitivity of the classification scheme used (Figure 4.12a). From the investigation of ChIP-Seq signal at single genes (Figure 4.12b), it was also evident that these newly classified PRC1-positive bivalent genes exhibit clear H2Aub1 and Ring1B signals using our own datasets, indicating that differences in the ChIP protocol may also account for differences in the classification.

We have found that the majority of bivalent genes is also H2Aub1-positive, and more importantly, H2Aub1 presence more closely represents bivalent chromatin than H3K27me3 since many genes solely marked by H3K27me3 (negative to H3K4me3 and H2Aub1) were found. This result suggests that PRC1 may have a direct role in restraining RNAPII elongation as previously suggested (Stock et al., 2007).

4.3.3 *RNAPII at some PRC-repressed genes is transcriptionally active*

To investigate the interplay between PRC function and RNAPII modification, we divided genes into PRC-positive (PRC^+, n=5628) and PRC-negative (PRC^-, n=9776).

At both PRC^+ and PRC^- genes, promoter levels of 8WG16 and S7P correlate with expression. S2P also increases with increasing expression, marking the PRC^+ genes with the highest activity levels. S5P in contrast, associates with the vast majority of PRC^+ promoters independently of expression; S5P enrichment at the end of genes parallels increasing expression at PRC^- genes (Figure 4.14). Surprisingly, the PRC^+ cohort contains not only silent genes, but also genes with intermediate and high expression, suggesting that PRCs do not act as blanket silencers (Schwartz and Pirrotta, 2008; Figure 4.13).

Different levels of H3K4me3, H3K27me3 and the correspondent correlation with actively transcribing RNAPII, could reflect a cell population with a proportion of cells in which the gene is either active or PRC-repressed. To further explore this possibility, in Chapter 5, we further classify the combinatorial patterns of RNAPII and PRCs using a clustering approach and test the co-existence of "active" PRC-targets by the use of sequential ChIP experiments, and single cell analysis of the correlation between chromatin modification and gene expression.

5

GENE ACTIVITY IN THE PRESENCE OF POLYCOMB

5.1 RESEARCH MOTIVATION

To investigate RNAPII regulation at PRC-targets genome-wide in ES cells, we mapped PRC subunits, PRC-instigated histone modifications and four RNAPII modifications by chromatin immunoprecipitation and high-throughput sequencing (ChIP-Seq). To characterise the different RNAPII regulation states at PRC-targets in ES cells, hierarchical clustering was used to group genes according to RNAPII modification and PRC occupancy. I aimed to investigate whether PRCs prevent the establishment of active RNAPII-S2P complexes or directly control their elongation. Moreover, I aimed to explore functional relevance of PRC-binding at "active" PRC-target genes and whether pluripotency factors might influence the switch between PRC-repression and the active state.

5.2 TWO ALTERNATIVE MODELS FOR POLYCOMB REGULATION AT "ACTIVE" PRC-TARGET GENES

The discovery of genes exhibiting enrichment for both PRCs and active RNAPII complexes and with high levels of mRNA led us to consider two alternative models of PRC-regulation at these genes, schematically depicted in Figure 5.1. Firstly, PRCs might regulate the level of transcription at any given gene by directly influencing elongation of active (S5P+S2P+) complexes, leading to an analog model of PRC regulation (Model 2). Secondly, PRC repression may not be compatible with the establishment of S5P+S2P+ complexes, in spite of elongation through coding regions by RNAPII complexes phosphorylated on S5 residues (Model 1). PRC-repressed genes exhibiting PRC-repressed and active RNAPII marks would therefore exist in

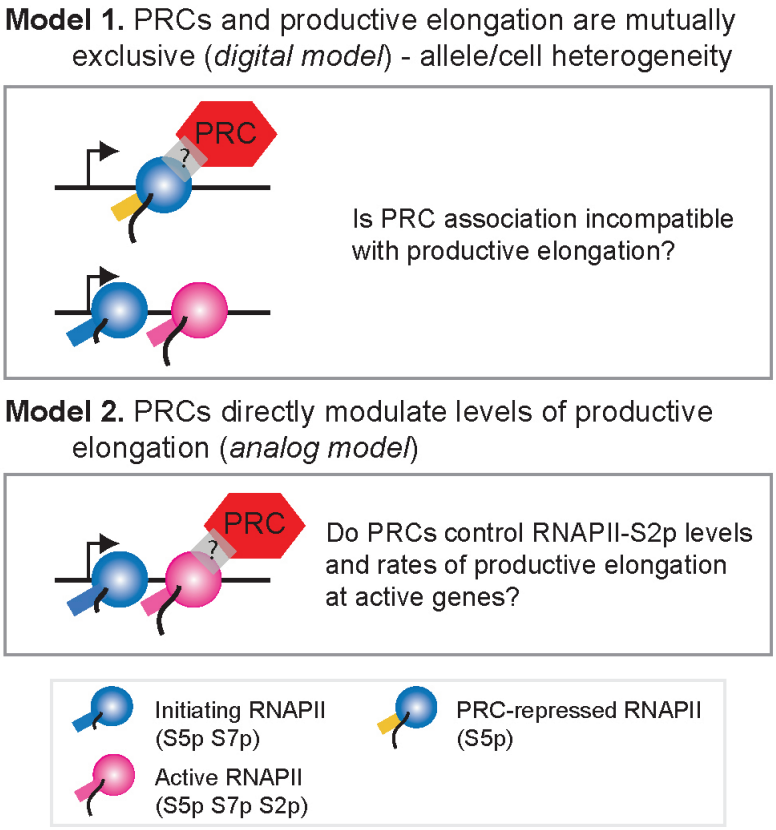


Figure 5.1: Two alternative models for PRC-regulation at PRCa genes.

different cells within the population (or different alleles within the same cell) due to variations in cell cycle or heterogeneity across cells. This results in an 'on-off' switch mechanism of regulation with two non-overlapping states, with S2P and PRCs binding in an exclusive manner, irrespectively of RNAPII-S5P association.

While true "poised" bivalent genes exist in mouse ES cells, it is equally possible that many PRC-repressed genes oscillate between an "active" and a "silent" state, based on productive transcription elongation. The latter situation can be considered a form of priming genes, since it indicates that these PRC-repressed genes are strongly prone to changes in the transcriptional factor network operating in ES cells. In this case, it remains to be explored which mechanisms could trigger the transition between the "active" and "PRC-repressed" state within the heterogeneous ES cell population.

5.3 RESULTS

5.3.1 *Identification of three epigenetic states in mESCs*

Having established that there is tight interplay between PRCs and RNAPII at PRC-repressed genes, I used hierarchical clustering to explore different RNAPII regulation states at PRC-targets in ES cells.

Average-link hierarchical clustering was used to analyse a matrix of 15404 non-overlapping RefSeq genes and 7 binary variables which consisted of ChIP-Seq dataset classified in Chapter 4, these were: S5P (TSS), 8WG16 (TSS), S7P (TSS), *S*5P (upstream TTS), S2P (downstream TTS), H3K27me3 (TSS) and H2Aub1 (TSS). The full hierarchical tree is shown in Figure 5.2.

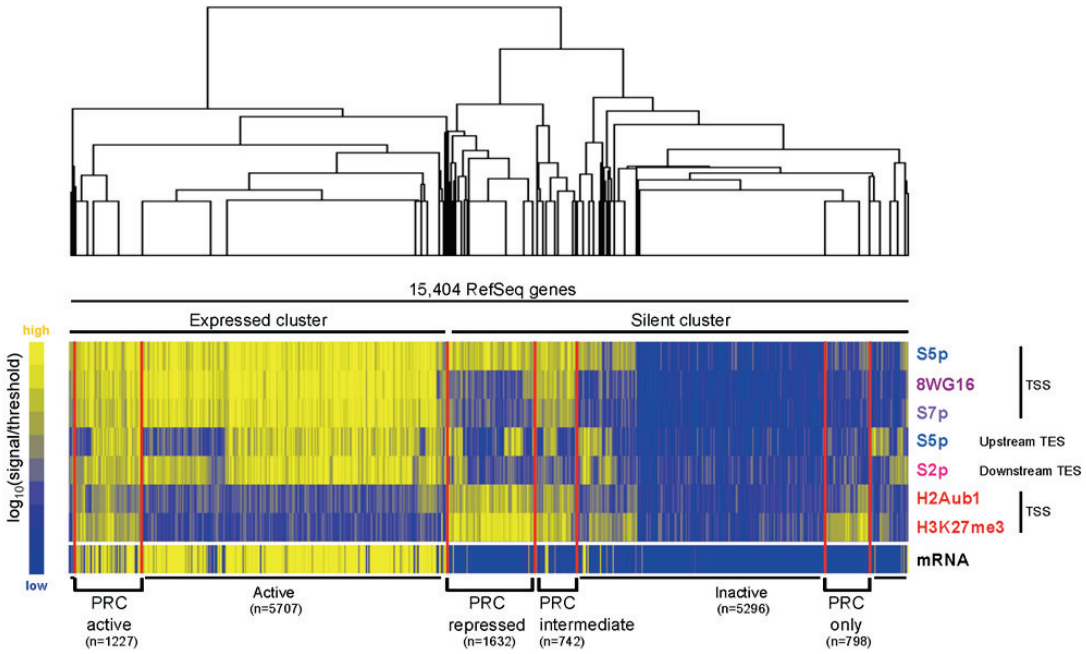


Figure 5.2: Hierarchical clustering of genes according to their combination of RNAPII and PRC instigated histone modifications. The patterns of RNAPII and PRC histone modifications across 15404 non-overlapping RefSeq genes were investigated by applying average linkage hierarchical clustering to a matrix consisting of seven variables. mRNA-seq levels for each gene are displayed, but were not used in the hierarchical clustering. The hierarchical tree shows that the genome subdivides first into 2 main clusters (broadly Silent and Expressed), and then further subdivisions appear, where specific PRC-target groups cluster. The $\log_{10}$(signal/threshold) of each variable was computed and divided into 10 percentiles, each assigned a color on a blue (low) to yellow (high) scale (left hand side). We extract four main PRC-bound groups: three from the Silent cluster and one from the Expressed.

An important step in hierarchical clustering is to select a distance measure, which will determine how the dissimilarity of two elements is calculated. For symmetric binary datasets, one can use the function *daisy* on R, available from the R package *cluster* (Maechler et al. 2005). Pairwise dissimilarities in the data matrix were computed using the Gower coefficient for symmetric binary datasets. The Gower coefficient is calculated as described in Gower [1971]: Given a 2 by 2 contingency table between two binary vectors i and j:

	i=1	i=0
j=1	a	b
j=0	c	d

The similarity *S(i,j)* and dissimilarity *D(i,j)* between i and j is given by:

$$S(i,j) = \frac{a+d}{a+b+c+d}$$

$$D(i,j) = 1 - S(i,j)$$

Clustering first divides genes broadly into three main groups (in order of appearance: "Active", "PRC-repressed" and "Inactive" groups; Figure 5.2). The "Active" and "Inactive" classes display similar characteristics to the previously classified 20% most expressed genes (Top20%) and 20% least expressed genes (Bottom20%), respectively (described in Chapter 3). The "Active" cluster is characterised by higher levels of RNAPII modification and the majority of genes within this branch do not recruit PRC, although a subclass of PRC-targeted genes is detected and was subsequently investigated. Genes at the "Inactive" branch of the tree, do not recruit RNAPII or PRCs and show little or absent levels of mRNA. Genes from the "PRC-repressed" branch, display characteristics reminiscent of the previously characterised "bivalent" genes, with high levels of PRC-mediated histone modifications (H3K27me3 and H2Aub1), and the presence of the initiating form of RNAPII (S5P) at all gene promoters classified within this branch. Some of these PRC-targeted genes display detectable S5P levels at the TTS consistent with elongation through coding regions.

5.3.2 *Principal component analysis identifies clear separation between Active, PRC-repressed and Inactive classes*

The described three classes ("Active", "PRC-repressed" and "Inactive") are also clearly visible when performing principal component analysis (PCA) using, instead of the binary classification, the continuous values (i.e. depth of sequencing at selected windows). PCA is a powerful tool for analysing data which involves a mathematical procedure that transforms multidimensional data in a way that maximizes the variability between objects. In other words, PCA identifies the most important gradients in the data. I performed PCA on a (N=15404) x (M= 7) matrix using *prcomp* function in R (R Development Core Team, 2010). Since PCA is most useful when variables are linearly related to each other, I log-transformed the data using log10 before performing the analysis. The structure of the data, consisting of three major concentrations/groups is clearly seen when plotting the first two principal components (PC1, PC2) against each other (Figure 5.3). Overlapping the first three classes of genes classified in hierarchical clustering ("Active", "PRC-repressed" and "Inactive") in the PCA plot shows a clear separation of the three groups by the principal components (Figure 5.3).

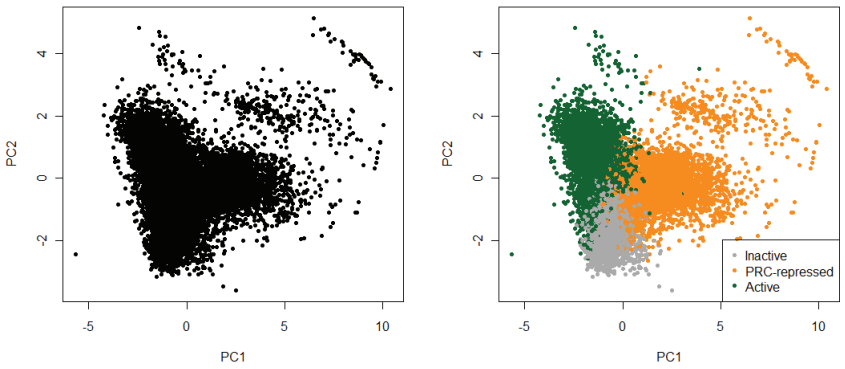


Figure 5.3: PCA identifies three major groups within the data according to their combination of RNAPII and PRC instigated histone modifications. PCA was applied to 15404 RefSeq genes described by 7 variables which included sequencing "depth of coverage" calculated at each TSS or TTS window: S5P (TSS), 8WG16 (TSS), S7P (TSS), S5P (upstream TTS), S2P (downstream TTS), H3K27me3 (TSS) and H2Aub1 (TSS). Plotting the first two principal components (PC1 and PC2) against each other shows three distinct groups of genes.

5.3.2.1 *How well do the components explain the variation in the data?*

The amount of variation in the data can be explained by each principal component. There are seven principal components (Table 5.1). The first component explains 63% of the variation in the entire data set, and the second and the third explain almost all of the remaining variation (18% and 7%, respectively). Components 1, 2 and 3 together explain 88% of the variance in the data.

Table 5.1: Variance within the data explained by the principal components. The first three principal components explain 88% of the variance and describe the principal gradients within the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Proportion of Variance	0.63	0.18	0.07	0.05	0.03	0.02	0.01
Cumulative Proportion	0.63	0.81	0.88	0.93	0.96	0.98	1

The loadings of each component can be used to understand which of the 7 variables contribute more strongly to each component (Table 5.2). One can interpret PC1 as being negatively related to marks of RNAPII. This component can be important to distinguish between "active" and "inactive" genes, where active genes have high negative scores and inactive genes have high positive PC1 scores. PC2, on the other hand, is highly related with the abundance of H2Aub1 and H3K27me3 signals, and is less sensitive to RNAPII. This component is important to distinguish between "poised" and "inactive"/ "active" genes. PC3 is highly correlated to levels of S5P at the 3' end of genes, but interestingly is negatively related to 8WG16 and S7P and practically unrelated to S5P at TSS region (Table 5.2).

5.3.3 *Comparing PCA with the results from Hierarchical clustering*

Comparing the principal components with the previously shown heatmap, we can see how well the components describe subtle variances within the clusters (Figure 5.4). The majority of genes in the Active group have a high negative contribution of PC1 and PC2, however a small percentage of genes within this group display high positive PC2; these genes correspond to genes with higher levels of H2Aub1 and H3K27me3 (highlighted in Figure 5.4). Similarly, a subgroup of genes

Table 5.2: Loadings of each principal component. PC1 is related to marks of RNAPII, PC2 is strongly correlated with PRC-mediated histone modifications (H2Aub1 and H3K27me3). PC3 is highly correlated to levels of S5P at the 3′ end of genes.

	*PC*1	*PC*2	*PC*3	*PC*4	*PC*5	*PC*6	*PC*7
S5P (TSS)	−0.43	0.03	−0.11	0.50	0.13	−0.72	0.13
8*WG*16	−0.44	0.19	−0.33	0.07	−0.02	0.22	−0.78
S7P	−0.43	0.22	−0.33	0.16	−0.13	0.52	0.60
S5P (TTS)	−0.37	0.13	0.85	0.16	0.19	0.24	−0.05
S2P (TTS)	−0.39	0.27	0.10	−0.74	−0.31	−0.33	0.09
*H2Aub*1	−0.32	−0.56	−0.15	−0.36	0.65	0.07	0.07
*H3K27me*3	−0.23	−0.71	0.11	0.12	−0.64	0.05	−0.06

within the Inactive cluster show a positive PC2 contribution contrasting to other genes in the Inactive group which have negative PC2 coefficients (highlighted in Figure 5.4).

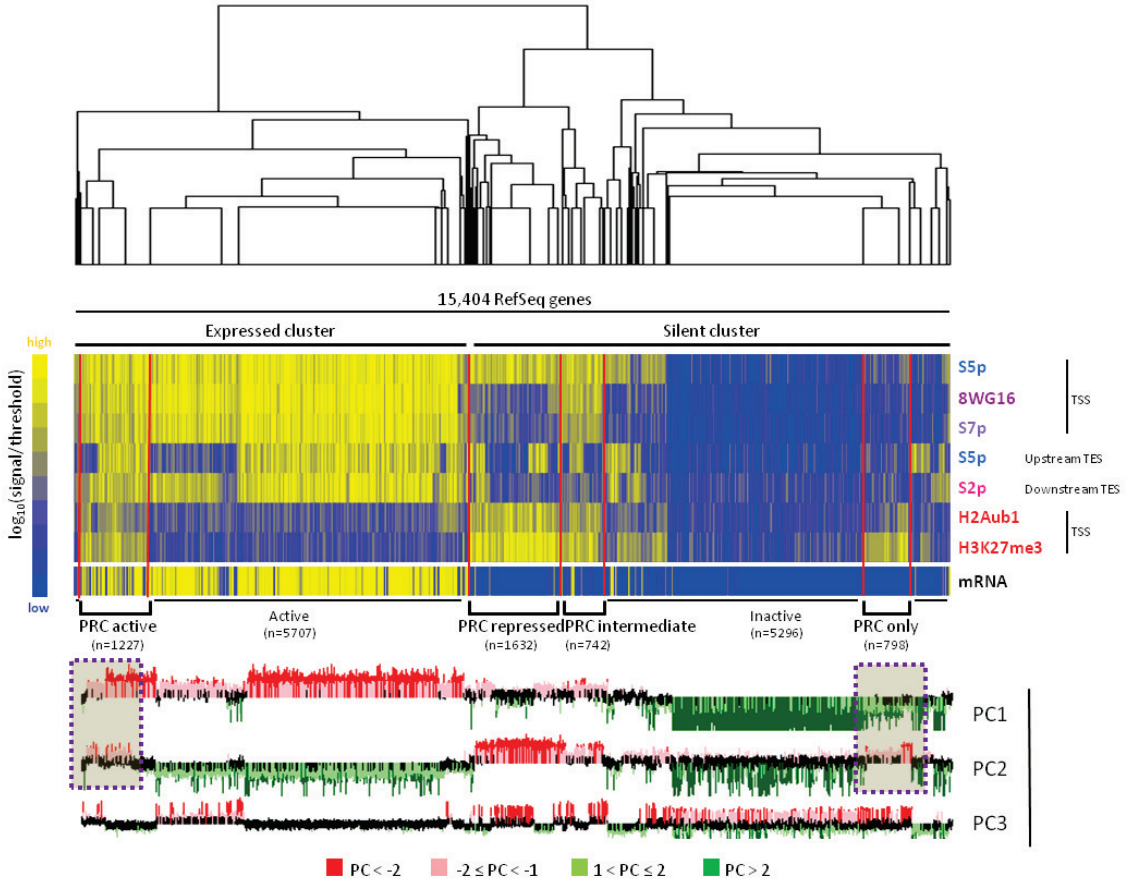


Figure 5.4: Comparison of Hierarchical clustering of genes with PCA results. PC1 distinguishes genes between "Active" and "Inactive" branches, PC2, on the other hand, is highly related with the abundance of H2Aub1 and H3K27me3. The combination of these two components reveals subtle variations within the clusters. A proportion of genes within the "Active" (gray shaded left box) and "Inactive" (gray shaded right box) branches have negative PC2 scores ; while the majority of "Active"/"Inactive" genes have positive PC2 scores.

5.3.4 *Subclassification of gene classes within the Active, PRC-repressed and Inactive branches of the HC tree*

We find that further biological relevant subdivisions appear further down the hierarchical clustering tree, distinguishable by presence/absence of RNAPII forms or PRC-mediated histone modifications (Figure 5.2; Figure 5.4). Within the "PRC-repressed" branch of the hierarchical tree, two separate subgroups of PRC-targets are characterised by presence of (i) H3K27me3, H2Aub1 and S5P,

without other RNAPII modifications (n=1632; PRCrepressed or PRCr), and (ii) H3K27me3, H2Aub1, S5P, 8WG16 and S7P, and little detectable S2P or expression (n=742; PRCintermediate or PRCi). Within the "Inactive" branch, a subgroup of PRC-target genes was also identified, this group was termed PRConly, or PRCo, and is characterised by the presence of H3K27me3, but little detectable H2Aub1 or RNAPII (n=798; PRConly or PRCo). Unexpectedly, we identified a fourth PRC-target cluster within the '"Active" branch, associated with all RNAPII modifications, H3K27me3, H2Aub1 and mRNA (n=1227; PRCactive or PRCa).

5.3.5 *Expression levels increase progressively from Inactive to Active groups*

Levels of mRNA increase progressively from Inactive, PRCo, PRCr, PRCi, PRCa to Active genes (Figure 5.5). High levels of mRNA are detected for the PRCi, PRCa and Active groups, low levels are detected for PRCr, while the majority of Inactive and PRCo genes have very low FPKM levels (57% and 47% of inactive and PRCo genes correspond to the 20% fraction of the least expressed genes in the genome, respectively). There is in good agreement between independent mRNA-seq analyses (present study, and Cloonan et al., 2008), while microarrays (Mikkelsen et al., 2007) are less sensitive to lower levels of expression. For example low mRNA expression from PRCr genes approaches background noise levels in microarray data, and thus becomes more complicated to detect visually and less clearly separated from the Inactive and PRCo classes (right plot, Figure 5.5), although statistically significant differences are observed (p-value = 5.3×10^{-12} and 1.83×10^{-12} when comparing PRCr expression levels with Inactive and PRCo groups, respectively in microarray data; Wilcoxon rank-sum test).

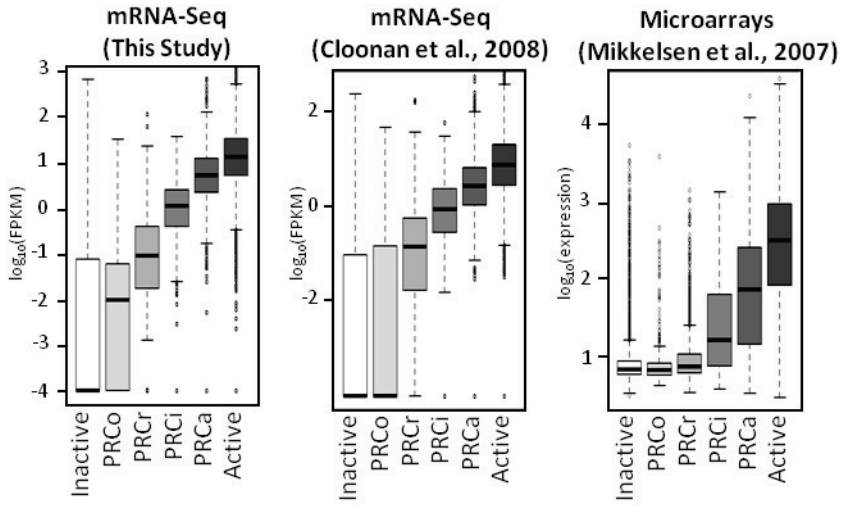


Figure 5.5: Levels of mRNA across Active, Inactive and PRC-cohorts. Expression levels were assessed using our own mRNA-Seq and published datasets (Sene et al., 2007, Mikkelsen et al., 2007, Cloonan et al., 2008). Active and PRCa display considerable amounts of mRNA-Seq levels, PRCi and PRCr display intermediate levels, while Inactive and PRCo display the lowest levels.

To confirm the expression status of distinct PRC gene clusters, we performed RT-PCR analysis on a panel of eight candidate PRCr genes (*Fgt5*, *Kcn4*, *Lrat*, *Lhx5*, *Pitx1*, *Zfp503*, *Gata5* and *Emx1*) and seven PRCa genes (*Rad50*, *Rrs1*, *Cldn5*, *Id3*, *Fzd8*, *Lefty2* and *Tbx3*). RT-PCR analysis involved the measurement of transcript levels from single genes and the use of α-amanitin, an inhibitor of RNAPII activity, which allows us to discern whether transcription is occurring, even if transcript levels are very low. Single gene RT-PCR analysis showed that transcript levels were detected at all analysed genes. Transcripts from PRCa genes are comparable to those of an active gene, and 100-1000 fold higher than PRCr genes (Figure 5.6). Transcripts are also sensitive to α-amanitin treatment (with exception of *Gata5*), thus confirming the transcriptionally active state of PRC-targets (PRCa).

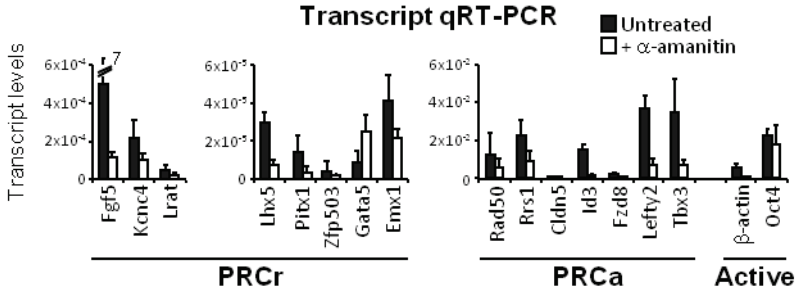


Figure 5.6: RT-PCR analyses confirm the expression status of genes from distinct PRC gene clusters. Levels of total RNA were measured by RT-PCR using primers located at the 5'-ends of genes, and were normalised to levels of housekeeping genes (*ß-actin*, *G6PD*, *Ubc*; exon-exon primers). Open bars show transcription level after α-amanitin treatment (7h, 75 µg/ml), which inhibits transcription. Transcripts are detected at all genes, shown by sensitivity to α-amanitin treatment, but levels are 100-1000x higher at PRCa/Active than PRCr genes, as expected. Mean and standard deviations are presented from 3 independent biological replicates. Experiment performed by E.Brookes.

5.3.6 *Characteristic chromatin signatures: From Inactive to Active groups*

To characterise the four PRC-target groups, we examined the levels of relevant markers within each of the six classes (Active, PRCa, PRCi, PRCr, PRCo and Inactive; Figure 5.7). A marked enrichment of S5P, 8WG16, S7P, H3K4me3 was observed within the promoters of Active and PRCa clusters, and to a lesser extent within the PRCi cluster. Consistent with productive elongation, S2P and H3K36me3 enrichments were observed at the 3' end of Active and PRCa clusters. These findings reveal that the class of PRCa genes correspond to a subset of PRC-targeted genes which acquire a chromatin signature comparable to that of transcriptionally active genes. On the other hand, at PRCr genes S5P extends throughout the coding regions without S2P and H3K36me3, and S5P enrichment is not detected, denoting that the abundance of RNAPII-S2P and H3K36me3 positively correlate with gene expression. We hypothesise that, analogous to what was discussed for bivalent genes, RNAPII termination at PRCr genes before transcription of polyadenylation signals, together with S2P absence, may limit mature mRNA production. At the other extreme, Inactive and PRCo groups display low levels of S5P and H3K4me3 and are depleted of RNAPII or histone marks of gene activity (H3K4me3 and H3K36me3).

The four PRC-cohorts (PRCo, PRCr, PRCi and PRCa) display the highest levels of H3K27me3 and H2Aub1, as expected. PRCr genes display the highest levels of occupancy of PRC subunits and associated histone marks; these levels decrease from PRCr to PRCi and PRCa genes (in parallel with the increase in mRNA levels). However, levels of H2Aub1, H3K27me3 and PRC subunits do not always inversely correlate with mRNA across the four PRC groups; in fact, the least expressed cohort, PRCo, has the lowest mRNA (Figure 5.5) and PRC levels (Figure 5.7). PRCo genes do not recruit RNAPII which is likely to explain the absence of mRNA detection. Below it is discussed how different promoter structures between PRCo and the other PRC cohorts (PRCr,i,a) can constitute the underlying influence for differences in the recruitment of PRC subunits and/or RNAPII to these promoters.

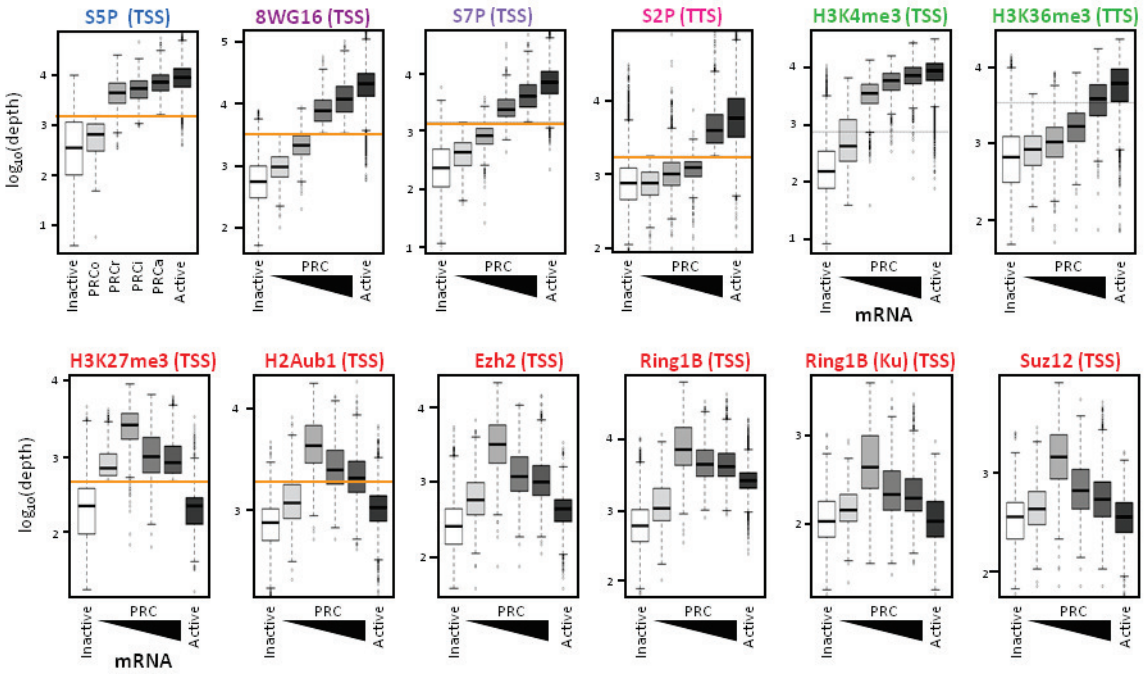


Figure 5.7: Levels of mRNA, RNAPII, and PRC subunits across Active, Inactive and PRC-cohorts. To investigate levels of RNAPII, PRC subunit occupancy and histone modifications, we assessed their enrichment (average depth/nucleotide) in 2 kb windows centered on TSSs or, for H3K36me3 and S2P, upstream or downstream of TTSs, respectively. PRC subunit occupancy is highest at PRCr, followed by PRCi, PRCa, PRCo, Active and finally Inactive. Solid orange line corresponds to the threshold applied after binary classification (classification results are detailed in Chapter 4 section 4.2.3). Ezh2, Ring1B and Suz12 (Ku et al., 2008), and H3K4me3 and H3K27me3 (Mikkelsen et al., 2007) refer to publicly available datasets.

5.3.7 *PRC-association is compatible with protein expression at PRCa genes*

PRC-association at PRCa genes was confirmed to be compatible with protein expression through mining of two ES cells proteome datasets (Van Hoof et al., 2006, Graumann et al., 2008). From the analysis of Graumann et al. [2008], peptides are positively identified from 32% and 15% of Active and PRCa genes, respectively, in comparison with 2-5% for other groups (Figure 5.8). This trend is also seen using data from another ES cell proteome study (Van Hoof et al., 2006; Figure 5.8). Amongst the PRCa genes expressed at the protein level are transcriptional regulators (*Hdac2*, *Hmga2*), cancer-linked genes (*Klf4*, *Kit*), and genes involved in glycolysis and pyruvate metabolism (*Hk1*, *Ldha*). This suggests that PRCs actively modulate expression levels of genes with important roles in ES cells, some of which are required for ES cell identity (e.g. *Hmga2* Hammond and Sharpless, 2008, *Klf4* Takahashi and Yamanaka, 2006) while others are involved in metabolic processes, such as glycolysis, that are differently regulated in ES and somatic cells (Kondoh et al., 2007). In support of our findings, several genes (*Hmg2a*, *Klf4*, amongst other such as *Tbx3*, *Fzd8*, *Lefty1*), here classified as PRCa, were also described as bivalent (Mikkelsen et al., 2007).

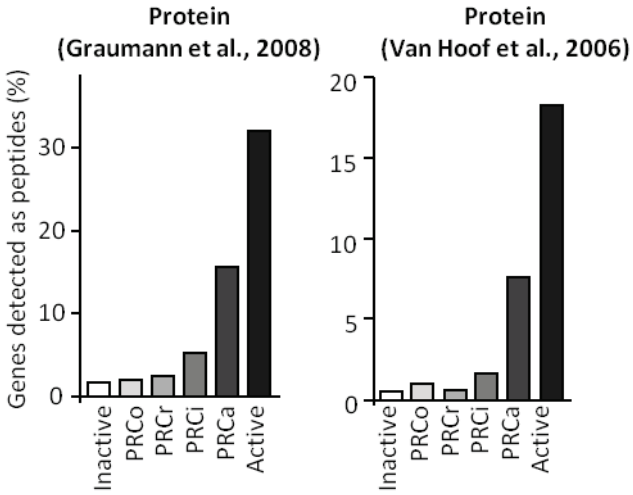


Figure 5.8: Gene expression at the protein level across Active, Inactive and PRC-cohorts. Number of genes within each cluster that are detected as peptides in ES cells according to two proteomic studies (Van Hoof et al., 2006, Graumann et al., 2008) were scored and expressed as a percentage of the total genes per group.

5.3.8 *Gene classes are distinguishable by promoter CpG content and DNA methylation*

To investigate the contribution of unmethylated CpGs to RNAPII and PRC recruitment across the different gene groups, I compared CpG content and DNA methylation data (Fouse et al., 2008). By measuring CpG percentage at each 2kb promoter (identified as ±1kb windows centered at each TSS), we verified that on average the Inactive and PRCo classes have the lowest CpG content, while the remaining PRC and Active promoters have high CpG content Figure 5.9.

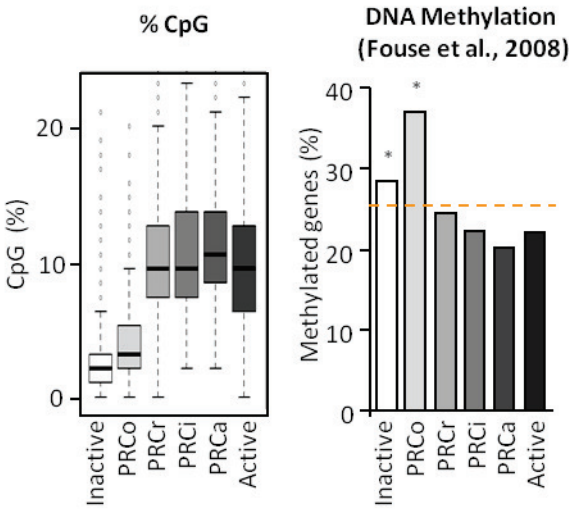


Figure 5.9: Percentage of CpG (%CpG) and DNA methylation at Active, Inactive and PRC cohorts. %CpG within 2 kb TSSs-centered windows is shown. High CpG PRC-cohorts are associated with high RNAPII-S5P occupancy. Genes with DNA methylation (Fouse et al., 2008) were significantly enriched in Inactive and PRCo cohorts (asterisk, p-value = $4.6\text{x}10^{-15}$ and $5.8\text{x}10^{-12}$, respectively; hypergeometric distribution). Dashed orange line represents proportion of methylated genes expected by chance.

I also investigated the proportion of promoters that overlap with CpG islands. CpG islands were downloaded from UCSC table browser (Karolchik et al., 2003), where they are defined according to Gardiner-Garden and Frommer [1987] (%GC > 50, CpG O/E > 0.6, length > 200bp), and overlapped with all gene promoters defined as 2kb windows centered at each TSS. The analysis revealed that PRCr, PRCi, PRCa and Active gene groups are particularly enriched for promoters overlapping CpG islands (p-value < 10^{-73}; hypergeometric distribution) while Inactive and PRCo promoters are depleted of CpG islands (p-value = 0.99; hypergeometric distribution). Several

studies have reported a strong association between high CpG promoter content and PRC recruitment (Tanay et al., 2007, Mendenhall et al., 2010). However, it was surprising to find that PRC-regulated genes (PRCo genes) had low frequency of CpGs, suggesting that PRC levels do not always reflect unmethylated CpG abundance and that alternative mechanisms may be employed to the recruitment of PRC2 to these genes.

Methylated genes are highly enriched in the Inactive and PRCo groups (p-value = $4.6x10^{-15}$ and $5.8x10^{-12}$, respectively; hypergeometric distribution; Figure 5.9). These two classes are also depleted of H3K4me3 and do not recruit RNAPII (Figure 5.7). The fact that methylated promoters are depleted of RNAPII and H3K4me3 has been demonstrated by several studies (reviewed in Cedar and Bergman, 2009). For instance, it has been suggested that when the preinitiation complex is loaded onto a given promoter, it might compete with DNA methyltransferases and their associated proteins mostly due to its huge size (Takeshima et al., 2009). Other studies have revealed that DNA-methylation and H3K4me3 are mutually exclusive, as promoter H3K4me3 marks prevent the association of DNMT3A/DNMT3L complex to the nucleosome (Ooi et al., 2007). The proportion of PRCo genes that are methylated (37%) are also modified by H3K27me3 to the average level of a gene within the PRCo group (p-value = 0.28 when comparing methylated PRCo genes with all PRCo genes, Wilcoxon rank-sum two-tailed test) suggesting that the two mechanisms of silencing, DNA methylation and Polycomb, might be coupled at these genes.

5.3.8.1 *Analysis of core H3 data*

To explore the possibility that lower levels of H3K27me3 and H2Aub1 in PRCo versus the other PRC groups (PRCr, PRCi, PRCa) are not due to lower core histone levels, I analysed publicly available core H3 ChIP-Seq data from mouse ES cells (Mikkelsen et al., 2007). ChIP-Seq data was downloaded as fastq files of raw sequenced reads from published datasets (GSM307624). Data processing was performed as described previously for other published datasets and as described in Chapter 2. Analysis of core H3 ChIP-Seq data shows nucleosome depletion at genes occupied by RNAPII (PRCr, PRCi, PRCa and Active) but not at genes from the PRCo

or Inactive groups (Figure 5.10). This result is in agreement with the current idea that the presence of the transcription complex leads to nucleosome exclusion in CpG-rich promoters, while CpG-poor promoters are in general more compact and less permissive to transcription initiation (Core et al., 2008, Juven-Gershon et al., 2008, Seila et al., 2008). This analysis also confirms that lower levels of PRC at PRCo and Inactive genes is not due to nucleosome depletion at their TSS.

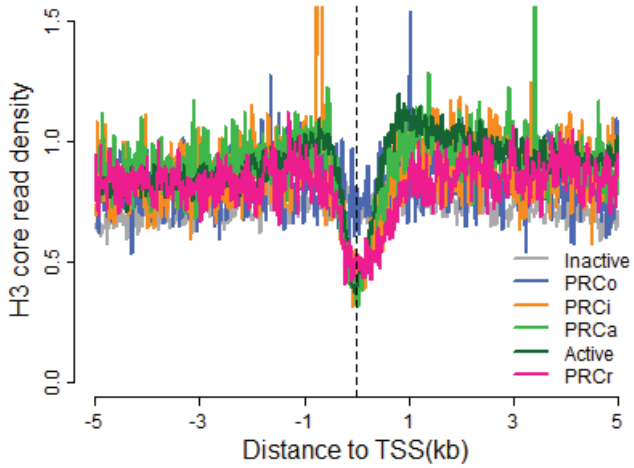
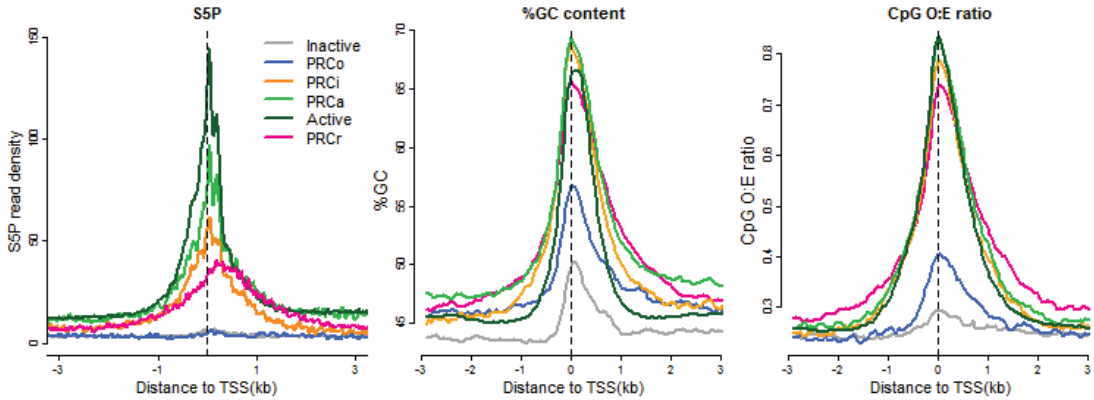


Figure 5.10: Genome-wide core H3 profiles at Active, Inactive and PRC cohorts. Average ChIP-Seq read density (depth per 10bp) was computed for the region surrounding the TSS of genes. Core H3 is particularly depleted from the TSS of gene groups associated with RNAPII-S5P (PRCr, PRCi, PRCa, Active), but not from genes not bound by RNAPII (Inactive, PRCo). This is consistent with competition between nucleosomes and RNAPII. The dashed line indicates the representative position of the TSS.

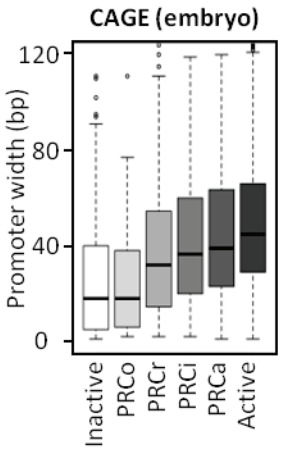
5.3.9 *PRCr gene promoters associated with broad S5P distribution and high CpG content*

Inspection of average TSS profiles reveals that PRCr genes have a S5P promoter peak that is unusually broad and delocalised by +190bp from the TSS (Figure 5.11a). The delocalised peak of the S5P profile may reflect a larger proportion of stalled RNAPIIs at the promoter region that could be explained by the slowing down of RNAPII or the usage of alternative TSS. Multiple TSS usage in broad promoter regions has been associated to broad unmethylated CpG-rich sequences (Sandelin et al., 2007) which in turn have been correlated with recruitment of PRCs (Tanay et al., 2007, Mendenhall et al., 2010) and RNAPII (Rozenberg et al., 2008). In fact, coincidentally with broader RNAPII-S5P distributions, we observed broader CpG densities around TSS of PRCr

genes than PRCa or Active genes (Figure 5.11a). To investigate the possibility of CpG rich promoters being over-represented by broad type of promoters which contain multiple transcription start positions (Carninci et al., 2006), we used CAGE data (Valen et al., 2009). Due to unavailability of CAGE data for mouse embryonic stem cells, the earliest developmental stage dataset available was used (E17.5 embryos). The analysis of CAGE data was performed by Dr. Boris Lenhard, Sars/CBU, Norway. CAGE data identifies "broad" promoter class at the gene groups associated with RNAPII and high CpG content (PRCr, PRCi, PRCa and Active) revealing that, as expected, high CpG density coincides with broad promoter usage in early development (Figure 5.11b). It is thought that nucleosome deficiency at RNAPII-bound genes is due to the binding of the the preinitiation complex which acts as a mechanism to prevent nucleosome formation and facilitate further gene activation (Gilchrist et al., 2010). Our analysis demonstrates that RNAPII recruitment is strongly associated with broad CpG-rich promoters of PRC-regulated genes (PRCr), emphasizing the fact that promoter structure of highly regulated genes and the recruitment of transcription machinery contribute significantly to precise control of gene expression (Carninci et al., 2006).



(a)



(b)

Figure 5.11: Correlation between broad CpG content and S5P profiles at Active, Inactive and PRC cohorts. (a) The average distribution of RNAPII-S5P and CpG was assessed at the six hierarchical clusters by analyzing the depth of sequencing reads and the CpG density around the TSS (±3 kb). Sequencing depth was calculated in 10 nucleotide windows. CpG O:E ratio and %GC were calculated in windows of length 200bp with an overlap of 150bp between windows. From PRCa to PRCr genes, the S5P promoter peak becomes broader and slightly shifted into the coding region. The distribution pattern of CpG O:E and %GC is also broader at PRCr than PRCa/Active genes. (b) Analysis of published CAGE data from E17.5 embryos (earliest developmental stage datasets available; Valen et al., 2009) shows that CpG density correlates with broad promoter usage in early development.

5.3.10 *Similar RNAPII and PRC enrichment patterns at PRCr genes*

To investigate the interplay between PRCs and RNAPII, I generated aggregate TSS profiles (Figure 5.12) for RNAPII S5P and PRC catalytic subunits (Ezh2 and Ring1B)

and mediated histone modifications (H3K27me3 and H2Aub1, respectively). PRCr genes show broader S5P enrichment profile then PRCi and PRCa in which the peak is sharper at the promoter. The broad region of S5P enrichment at PRCr genes is accompanied by broad peaks of H3K27me3 and H2Aub1. Profiles of Ezh2 and Ring1B show very similar patterns to the histone modifications they catalyse and RNAPII-S5P.

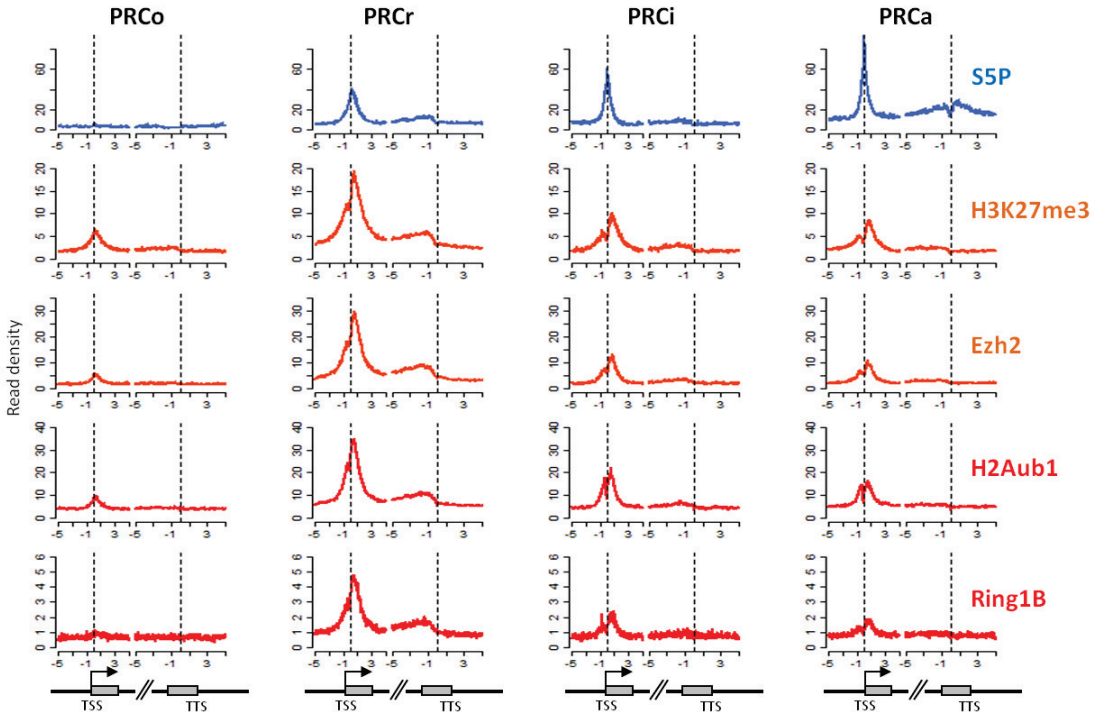


Figure 5.12: S5P and PRCs show similar genome-wide ChIP profiles at PRCr genes. Average ChIP-Seq read density (depth per 10bp) was computed for the region surrounding the TSS and TTS of PRCo, PRCr, PRCi and PRCa genes. While S5P profile at PRCi and PRCa genes resembles the one obtained for highly expressed genes, the profile at PRCr genes shows a much broader domain of enrichment which is accompanied by the domain of enrichment of PRC subunits (Ezh2, Ring1B) and mediated histone modifications (H3K27me3, H2Aub1), thus suggesting a interplay between RNAPII and PRCs.

From average profiles, it is notable that there is a small enrichment of S5P at the 3′ end of PRCr genes, consistent with RNAPII elongation up to the 3′ end. To confirm that PRCs and RNAPII are detected throughout coding regions of PRCr genes, ChIP-Seq profiles at genes classified as positive or negative for S5P immediately upstream TTS were investigated (Figure 5.13). We verified that at some genes, S5P is limited to the promoter region (*Kcn4*, *Emx1*), while others (*Foxa2*, *Lbxcor1*) display

larger domains of enrichment although, unlike at Active or PRCa genes, without detectable accumulation of S5P downstream the TTS was observed. This revealed that the domain of S5P enrichment is similar to the one of H2Aub1 and H3K27me3 thus suggesting an interplay between the two complexes at PRCr genes.

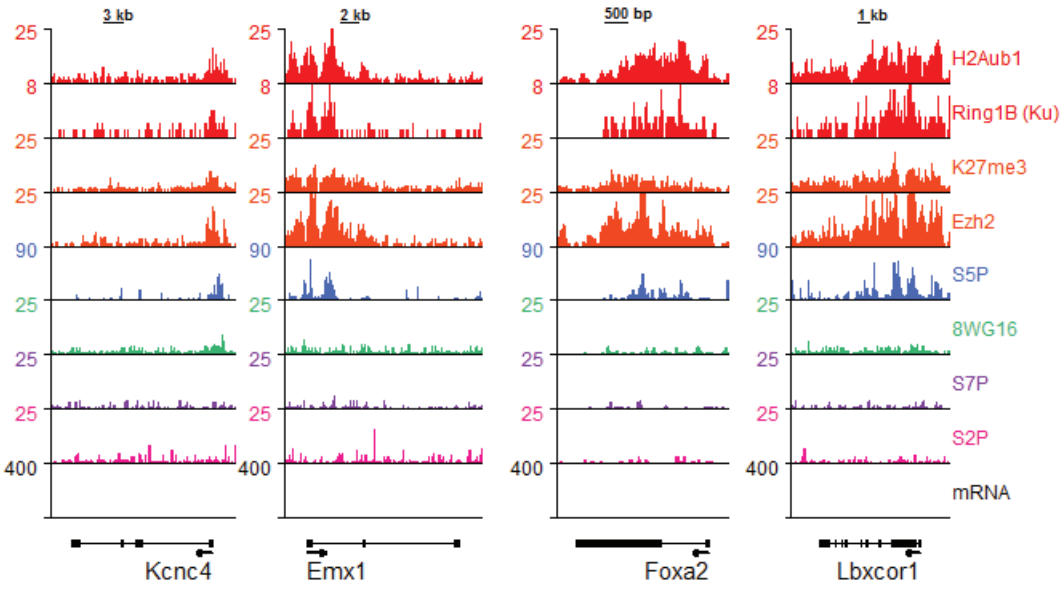


Figure 5.13: Similar domains of enrichment of S5P, H3K27me3 and H2Aub1 at PRCr genes. ChIP-Seq profiles are shown at representative PRCr genes classified as negative (*Kcnc4*, *Emx1*) and positive (*Foxa2*, *Lbxcor1*) for S5P in the TTS (2kb upstream). Y-axes represent the number of reads. S5P, H3K27me3 and H2Aub1 show variable degrees of extension through gene bodies. Very low levels of 8WG16, S7P, S2P and mRNA are detected. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

5.3.11 *PRC1, PRC2 and RNAPII physically co-associate at PRCr genes*

Remarkably similar S5P and PRC distributions at TSSs and throughout coding regions of PRCr genes suggest that PRCs travel through, and dissociate from, genes coincidental with elongating S5P complexes. In order to investigate whether RNAPII and PRCs co-associate at PRC genes, we performed sequential ChIP (re-ChIP). The co-occupancy between S5P and either Ring1B or Ezh2 was analysed at six PRCr genes (Figure 5.14): three with S5P and Ring1B more localised at promoters (*Lhx5*, *Pitx1*, *Zfp503*) and three showing coding region enrichment (*Fgf5*, *Gata5*, *Lrat*).

Additionally, we performed S5P/Ring1B re-ChIP experiments at eight previously characterised PRC-repressed genes (Stock et al., 2007; data not shown).

Prior to sequential ChIP experiments, the enrichments of S5P, Ezh2 and Ring1B were confirmed by qPCR as shown in (Figure 5.14, left panel). High promoter levels of S5P expected from ChIP-Seq studies were confirmed for all PRCr genes (*Lhx5*, *Pitx1*, *Zfp503*, *Fgf5*, *Gata5*, *Lrat*), and the presence of RNAPII at the TTS was also confirmed for the genes which scored positive for the association of this mark to the 3′ end region (*Lhx5*, *Pitx1*, *Zfp503*). Re-ChIP experiments detect enrichment for RNAPII-S5P with Ring1B or Ezh2, independently of immunoprecipitation order demonstrating that RNAPII and PRC proteins bind to the same chromatin at PRCr genes (Figure 5.14, right panel). At the 3′ end of PRCr genes, co-association between PRC1, PRC2 and S5P is seen for genes in which RNAPII was detected throughout the gene's body by ChIP-Seq profiles (*Lhx5*, *Pitx1*, *Zfp503*), but not for those where the marks are more localised at the promoters (*Fgf5*, *Gata5*, *Lrat*), confirming the quality of ChIP-Seq datasets. Control experiments included Ring1B followed by Ring1B (positive) and Ring1B followed by mock immunoprecipitation (negative), the latter demonstrating low level antibody carry-over to the second immunoprecipitation.

These results demonstrate that S5P, PRC1 and PRC2 co-occupy promoters of PRCr genes, and together extend by a variable distance into coding regions. These detailed single-gene analyses, together with similar distributions of RNAPII-S5P and PRCs through PRCr genes in genome-wide mapping, demonstrate co-occupancy of RNAPII-S5P and PRCs at TSS and coding regions in ES cells.

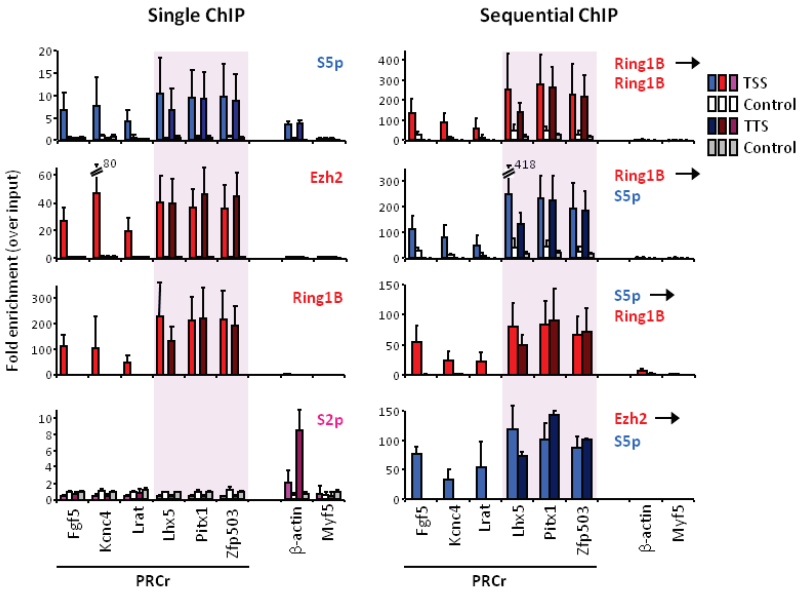


Figure 5.14: PRC1, PRC2 and RNAPII physically co-associate at PRCr genes. Ring1B, Ezh2 and S5P enrichments analysed by ChIP qPCR and re-ChIP at Active (*ß-actin*), Inactive (*Myf5*), PRCr genes. PRCr genes are subdivided into those with (*Lhx5*, *Pitx1*, *Zfp503*) or without (*Fgf5*, *Gata5*, *Lrat*) detectable S5P enrichment upstream of TTSs. Enrichment expressed relative to input DNA using the same DNA amount in the PCR. Background levels (mean enrichment from control antibodies and beads alone) at TSS (white bars) and TTS (grey bars). Mean and standard deviations from 3-6 independent biological replicates. Ring1B, Ezh2, S5P and S2P are found at expected levels. Sequential ChIP analysis shows that S5P co-associates with Ring1B and Ezh2 at the same chromatin in all cases. No DNA was immunoprecipitated from S5P –> mock or Ezh2–> mock, and so control bars are not shown for S5P –> Ring1B or Ezh2 –> S5P.

5.3.12 *Ring1B depletion affects the expression levels of PRCr, PRCi and PRCa genes*

To investigate the changes in gene transcription that occur at PRC-target genes following Ring1B depletion, published microarray expression data was analysed for Ring1A-knockout cells in which a conditional Ring1B-knockout is induced by tamoxifen (Endoh et al., 2008). PRCr, PRCi and PRCa genes show substantial changes in expression and were predominantly up-regulated after Ring1B-knockout (29%, 32% and 30% of genes, respectively, showed a >1.5 fold increase in expression) which is in agreement with Ring1B acting as a transcriptional repressor (Figure 5.15a). The mean expression change at PRCa genes after Ring1B-knockout is lower than at PRCr genes (p-value=0.022, one-tailed Wilcoxon rank-sum test), likely due to the fact that

transcript levels of PRCa genes are higher then PRCr genes in wild-type ES cells. PRC1 repression was further confirmed by qPCR analysis over a Ring1B-knockout time-course on a selection of twelve genes (Figure 5.15b) these included: six PRCr genes (*Fgf5*, *Kcnc4*, *Lrat*, *Lhx5*, *Pitx1*, *Zfp503*) and six PRCa genes (*Fzd8*, *Lefty2*, *Tbx3*, *Hmga2*, *S100a6*, *Tpm4*). Interestingly, qPCR results suggested that de-repression is greater for genes with S5P and PRC marks extending through gene coding regions (*Lhx5*, *Pitx1*, *Zfp503*), than for genes where both activities are more restricted to the TSS (*Fgf5*, *Kcnc4*, *Lrat*; Figure 5.15b).

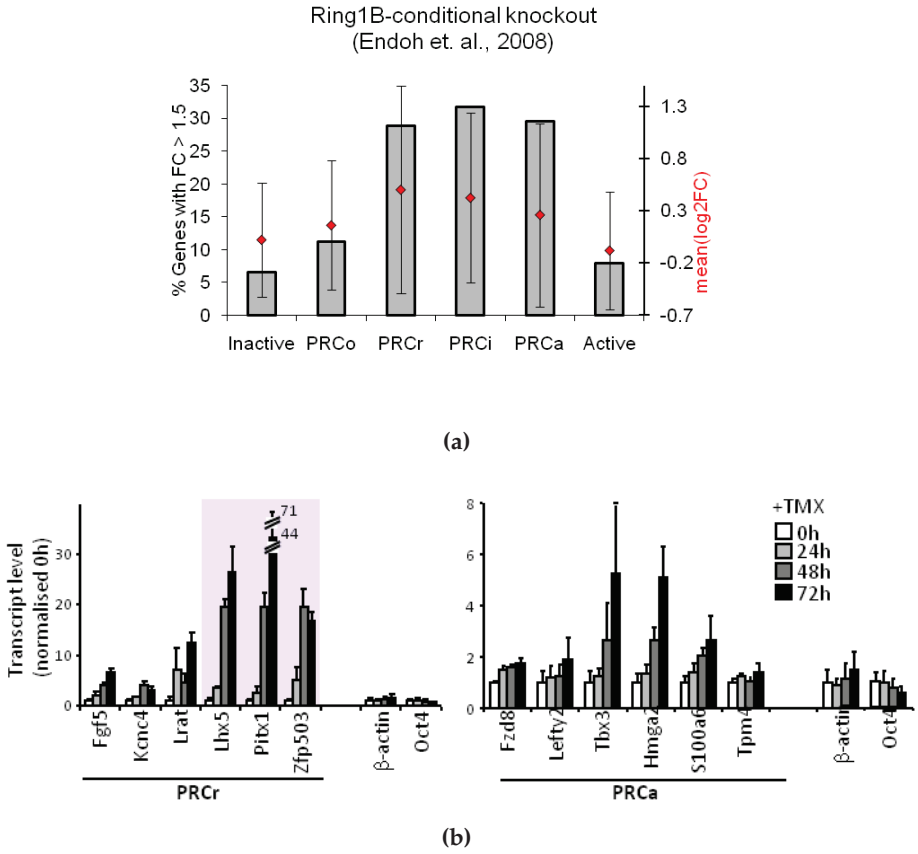


Figure 5.15: PRCr, PRCi and PRCa genes are predominantly up-regulated after Ring1B-knockout. (a) Analysis of published microarray data (Endoh et al., 2008) for ES-ERT2 cells before and after 48h tamoxifen (TMX) treatment. Barplot represents the percentage of genes which become up-regulated (i.e. showing >1.5-fold increase) after Ring1B removal. The percentage of up-regulated genes is statistically significant for PRCa, PRCi and PRCr cohorts (p-value < 10^{-36}; hypergeometric test). Red marks represent the mean $\log_2$ fold change, and vertical error bars represent the standard deviation. (b) RNA levels measured in ES-ERT2 Ring1A-knockout cells in which Ring1B is removed by tamoxifen addition. Transcript levels were assessed by qPCR using primers located at gene 5′ ends, and normalised to housekeeping genes (*ß-actin*, *G6PD*, *UBC*), and then to 0h treatment. PRC-repressed genes become de-repressed upon Ring1B removal; the effect is more marked at genes where S5P extends up to the TTS. Mean and standard deviations presented from three independent biological replicates. Experiments were performed by E. Brookes.

Analysis of genome-wide data supports this conclusion in respect to the PRCr group (Figure 5.16), with genes classified as positive for S5P at the TTS (S5PEnd+) being more likely to be de-repressed and show higher change in expression levels than those where S5P is not at the TTS (S5PEnd-; p-value = 0.0021; one-tailed Wilcoxon rank-sum test). To exclude the possibility that S5P detection at the TTS is due to

shorter gene length (and thus promoter-proximal S5P occupancy is detected in close proximity of the TTS), genes which are less then 5kb long were excluded from the analysis, and the results obtained were confirmed even when excluding smaller genes (Figure 5.16). These results suggest that the presence of RNAPII-S5P in coding regions of PRC-repressed genes, accompanied by proportional PRC occupancy, may favor gene activation upon PRC removal.

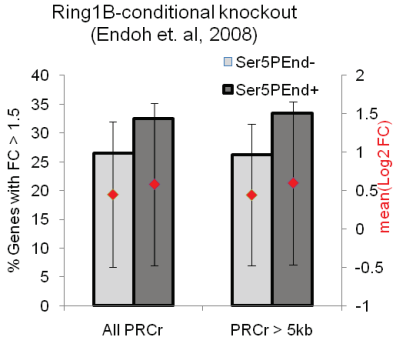


Figure 5.16: Genes classified as positive for S5P at the TTS (S5PEnd+) are more likely to be de-repressed after Ring1B removal. The number of up-regulated genes and the mean fold change in expression is higher for PRCr with S5P and PRCs at both TSS and TTS. Barplot represents the percentage of genes which become up-regulated after Ring1B removal. Red marks represent the mean $\log_2$ fold change, and vertical error bars represent the standard deviation. After removing genes shorter then 5kb, the same trend still applies.

5.3.13 *PRCs functionally repress active developmental and metabolic genes in ES cells.*

To determine if the six groups of genes are associated with particular biological processes, GO and KEGG pathway analysis were performed. GO analysis revealed that general cellular functions such as “ribosome”, “DNA repair” and “nucleic acid metabolic process” are predominantly found in the “Active” group, whereas genes in the “Inactive” and “PRCo” groups are associated to more specific processes such as “receptor binding”, “response to stimuli”, “signaling”, “‘multicellular organismal process” and “signal transduction”. While PRC-targets lacking RNAPII recruitment (PRCo) are associated with adult-specific processes where expression is required in terminally differentiated cells, PRCr, PRCi and PRCa genes are specifically enriched for genes associated with developmental processes such as genes from the *Fox*, *Hox* and *Sox* families. Furthermore, despite the association to developmental processes,

PRCa genes are also enriched for metabolic GO terms (p-value < 10^{-15}; hypergeometric distribution).

KEGG pathway analysis identify PRCa genes associated with TGFß-, Wnt- and MAPK-signaling pathways, and with cancer, cell cycle, and energy metabolism (Figure 5.17). Although de-regulation of members of the TGFß-signaling pathway and other cellular communication pathways has been reported after conditional knockout of Ring1B (Van Der Stoop et al., 2008), to our knowledge no direct regulation of metabolic genes by PRCs in ES cells has been shown or explored before. Summarizing, these results suggest that PRC might be modulating the levels of metabolic enzymes, in addition to the cited roles on repressing signaling and developmental genes in ES cells.

KEGG pathway *(p-value)*	Gene Symbols *(PRCa gene members)*
Pathways in cancer *($2x10^{-6}$)*	*Stat1; Gli2; Ralb; Rassf5; Hdac2; Bcr; Fgf18; Tcf7; Stat5b; Lamb1; Egln3; Traf3; Akt1; Fgf17; Myc; Pdgfb; Cdkn1a; Vegfa; Lama1; Epas1; Fzd8; Vegfb; Rxra; Plcg1; Nfkb1; Tgfbr1; Cdkn2b; Cdk6;Kit; Fzd10; Pdgfa; Smo; Tcf7l1; Tgfb1; Igf1r; Fgf3; Ccnd1; Fgfr1; Pik3r2; Mmp2; Plcg2; Smad3*
TGF-beta signaling pathway *($4x10^{-6}$)*	*Bmpr2; Inhbb; Lefty2; Lefty1; Id2; Id4; Myc; Id1; Bmp7; Tgfbr1; Cdkn2b; Bmp8a; Id3; Tgfb1; Tfdp1; Smad3; Smad6*
MAPK signaling pathway *($1x10^{-5}$)*	*Mapkapk2; Gadd45b; Fgf18; Dusp14; Cacnb1; Dusp3; Map3k14; Map3k9; Akt1; Gadd45g; Flnb; Nfatc4; Fgf17; Myc; Pdgfb; Mapk12; Mapk11; Mapk8ip2; Mapk4; Rasgrp2; Rps6ka4; Dusp5; B230120H23Rik; Cdc25b; Pla2g12a; Nfkb1; Mos; Tgfbr1; Pdgfa; Tgfb1; Fgf3; Fgfr1; Dusp4; Rasa2; Mras; Cacna2d2; Dusp9*
Cell cycle *($1x10^{-3}$)*	*Hdac2; Gadd45b; Gadd45g; Cdc14b; Myc; Smc1b; Cdkn1a; Cdc25b; Cdkn2b; Sfn; Cdk6; Tgfb1; Ccnd1; Tfdp1;Smad3*
Wnt signaling pathway *($1x10^{-3}$)*	*Vangl2; Camk2b; Tcf7; Ppp2r5c; Sfrp4; Nkd2; Nfatc4; Myc; Fzd8; Lrp5; Fosl1; Frat1; Sfrp2; Vangl1; Fzd10; Tcf7l1; Ccnd1; Sfrp1; Smad3*
p53 signaling pathway *($3x10^{-3}$)*	*Steap3; Gadd45b; Igfbp3; Gadd45g; Cdkn1a; Pmaip1; Zmat3; Sfn; Cdk6; Ccnd1*
Inositol phosphate metabolism *($6x10^{-3}$)*	*Synj2; Pip5k1b; Pip4k2a; Pip5kl1; Plcg1; Pik3c2a; Plcg2*
Pyruvate metabolism *(0.01)*	*Pck2; Acss1; Akr1b3; Ldhb; Ldha; Dlat*
Notch signaling pathway *(0.01)*	*Hdac2; Jag1; Hes5; Lfng; Dll3; Aph1c*
ErbB signaling pathway *(0.02)*	*Camk2b; Stat5b; Akt1; Myc; Cdkn1a; Hbegf; Plcg1; Pik3r2; Gab1; Plcg2*
Citrate cycle (TCA cycle) *(0.04)*	*Ogdh; Pck2; Dlat*
Glycolysis / Gluconeogenesis *(0.04)*	*Hk1; Pck2; Acss1; Eno2; Ldhb; Ldha; Dlat*

Figure 5.17: . KEGG pathways significantly enriched in active PRC-target genes (PRCa cluster). Enrichment of KEGG pathways by group members was assessed by hypergeometric testing.

We were interested in the evidence for metabolic control of pluripotency. To investigate the functional importance of the expression of active PRC target genes, we analysed temporal changes in gene transcription of PRCa metabolic genes (485 genes associated to GO "metabolic process" GO:0008152) following LIF withdrawal o investigate their regulation in ES cells. The analysis of published microarray expression data (Shen et al., 2009) revealed that many metabolic PRCa genes are down-regulated during ES cell differentiation (Figure 5.18), including genes involved in glycolysis and pyruvate metabolism (*Ldha*, *Gpd1l* and *Pck2*; Figure 5.17). Other metabolic genes become up-regulated during differentiation (Figure 5.18), as is most common for developmental regulator genes, suggesting roles during lineage

specification. Regulation of PRC-targets metabolic genes during ES cell differentiation extends our understanding of Polycomb function in pluripotency to include the modulation of ES cell metabolism, in addition to repression of developmental genes.

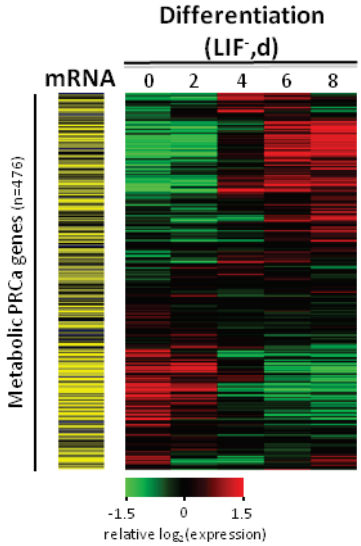


Figure 5.18: Changes in expression during differentiation for PRCa genes. Analysis of microarray expression data for ES cells differentiation after LIF withdrawal (Shen et al., 2009) shows that metabolic PRCa genes can become up- or down-regulated upon differentiation. The median normalised log intensities for a total of 476 PRCa genes (out of 485) represented in the array and annotated with "metabolic process" Gene Ontology are represented (GO:0008152). Genes were ordered according to hierarchical clustering performed using MEV (Saeed et al., 2003). Left yellow-blue bar represents FPKM levels coloured relative to all RefSeq genes; yellow = expressed and blue = repressed. Microarray analysis of raw datasets were performed by T.Carroll.

5.3.14 *PRCs physically co-associate with RNAPII-S5P but not S2P at active PRC-targets*

The detection of PRC-mediated histone modifications in combination with active RNAPII modifications at PRCi and PRCa genes was confirmed by the inspection of average TSS and TTS profiles of the four classes of PRC-targets (Figure 5.19). Levels of activating RNAPII modifications increase from PRCo to PRCa genes. At PRCa genes, enrichment profiles of S5P, 8WG16, S7P and S2P resembles those found for active genes but on average at lower levels, with peaks of S5P at the promoter and downstream of the TTS, together with S2P enrichment peak after the TTS, despite the presence of the repressive marks H3K27me3 and H2Aub1.

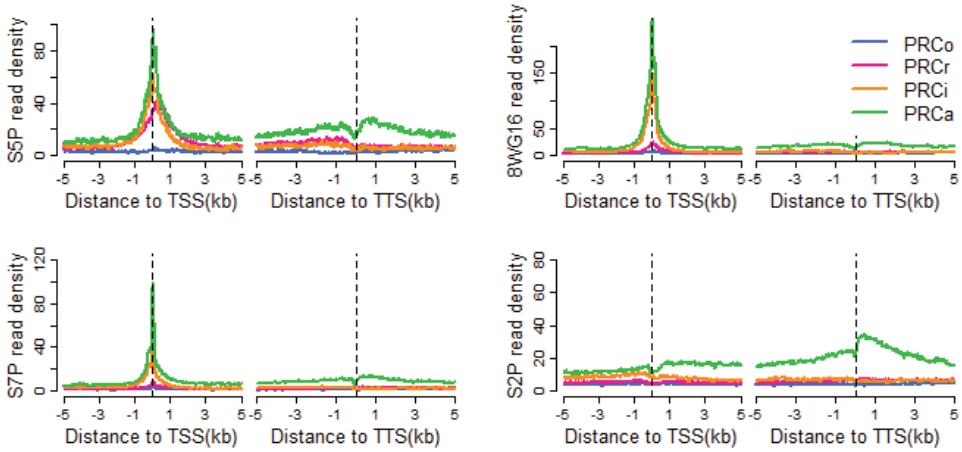


Figure 5.19: Distinct combinations of RNAPII at each PRC class. Average ChIP-Seq read density (depth per 10bp) profiles were computed for the region surrounding the TSS and TTS of PRCo, PRCr, PRCi and PRCa genes. PRCo genes do not recruit RNAPII and thus noise RNAPII levels are seen for this class. PRCr, PRCi and PRCa show a peak or RNAPII-S5P at the promoter region. 8WG16 and S7P levels increase from PRCi to the PRCa class. Furthermore, PRCa genes are also associated with S2P.

The different combinations of RNAPII and PRCs at PRCr, PRCi and PRCa genes was further confirmed by the inspection of single gene profiles (Figure 5.20) and ChIP qPCR (Figure 5.21). Figure 5.20 shows selected genes from the PRCi (*Atf3*, *Crabp1*) and PRCa (*Id3*, *Tbx3*). At *Aft3* and *Nppc* PRCi genes, H3K27me3 and H2Aub1 marks are accompanied by peaks of 8WG16 and S5P localised at the promoters and a S7P peak at *Nppc* promoter, with low S2P levels and no mRNA. At *Id3* and *Tbx3* PRCa genes, broad enrichment of S5P, 8WG16 and S7P over the gene's body is detected, and remarkable levels of S2P and mRNA are seen, despite the presence of H3K27me3 and H2Aub1.

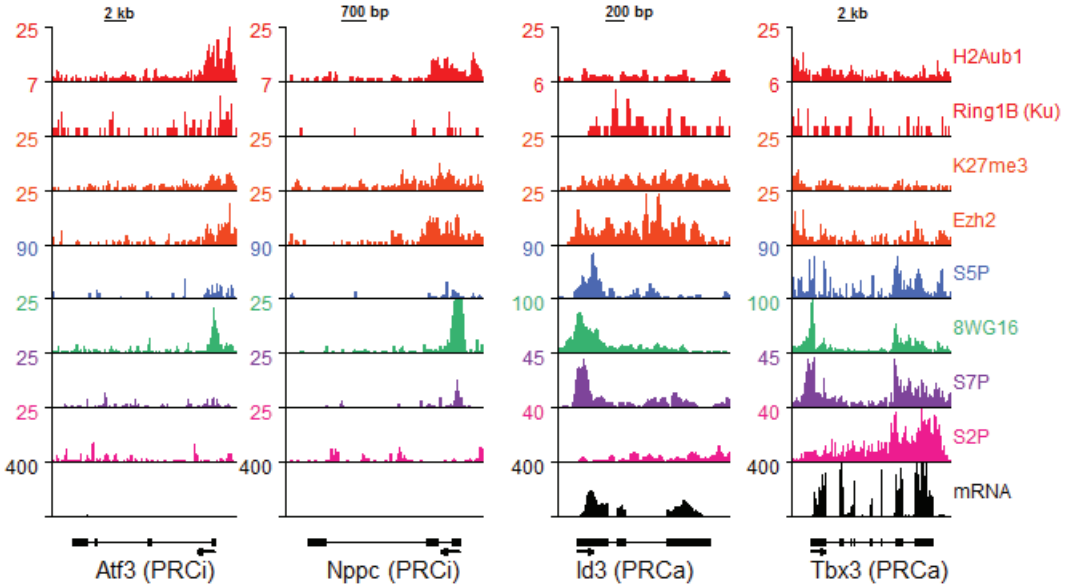


Figure 5.20: Simultaneous detection of repressive PRCs and activating RNAPII modifications at PRCi and PRCa genes. ChIP-Seq profiles are shown at representative PRCi (*Atf3*, *Nppc*) and PRCa (*Id3*, *Tbx3*) genes. All genes display considerable levels of RNAPII-S5P and the repressive marks H3K27me3 and H2Aub1. PRCi genes have low levels 8WG16 and S7P and no detectable S2P. At PRCa genes, despite the presence of repressive histone modifications, remarkable levels of 8WG16, S7P, S2P and mRNA are detected. Y-axes represent the number of reads. Arrows indicate gene start and direction of transcription. Black boxes indicate exons.

Simultaneous detection of repressive PRCs and activating RNAPII modifications at PRCi and PRCa genes by our ChIP-Seq mapping, led us to investigate whether PRCs prevent the establishment of active RNAPII-S2P complexes (Figure 5.1, Model 1 or directly control their elongation at PRCa genes (Figure 5.1, Model 2). In Model 1, PRCs would re-ChIP with RNAPII-S5P but not with RNAPII-S2P, and the co-existence of PRC-repression and RNAPII-S2P (and productive transcription) could be explained by separated chromatin states either in different alleles or in the heterogeneous ES cell population, while in Model 2 PRCs would re-ChIP with both S5P and S2P.

To investigate the co-association of PRCs with active (S2P) RNAPII, we have performed re-ChIP experiments with Ring1B and S2P of six PRCa genes (*Cldn5*, *Rad50*, *Rrs1*, *Id3*, *Left2*, *Tbx3*; Figure 5.21), three PRCi genes (*Atf3*, *Crabp1*, *Nppc*; not shown) and six PRCr genes (*Fgf5*, *Gata5*, *Lrat*, *Lhx5*, *Pitx1*, *Zfp503*, not shown).

Re-ChIP of S5P with S2P confirms the presence of S5P with S2P at active, PRCi and PRCa genes (Figure 5.21), but not PRCr. However, consistent with Model 1, re-ChIP experiments demonstrate that Ring1B does not co-localise with S2P, despite positive co-occupancy with S5P (Figure 5.21). Thus, lack of co-association between PRCs and RNAPII-S2P, together with genome-wide analyses, suggests an 'on-off' switch mechanism of PRCa genes where RNAPII-S5P exists in two independent states: (a) in association with PRC (in the absence of other active marks) and (b) in the presence of S2P (and absence of PRCs), which may be due to ES cell heterogeneity or allelic exclusion. Variable expression status across the cell population may result from heterogeneity, which has been previously described for ES cells (reviewed in Graf and Stadtfeld, 2008).

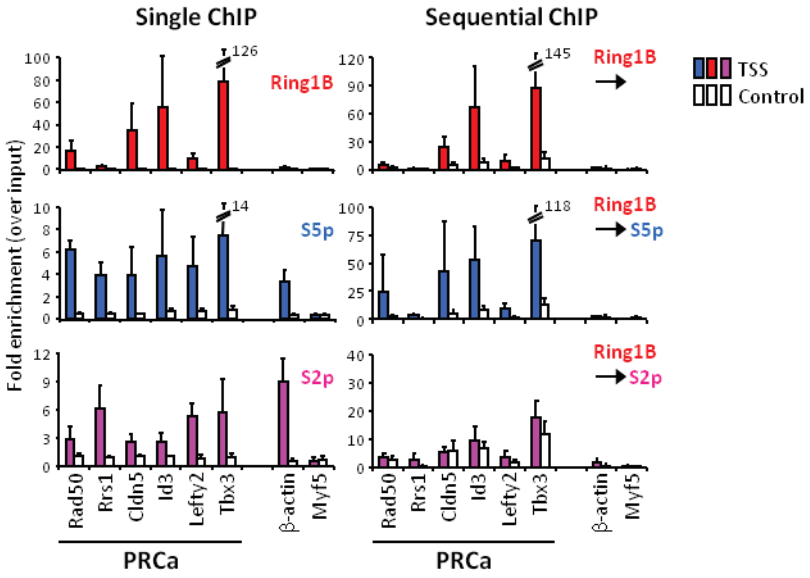


Figure 5.21: PRCs physically co-associate with RNAPII-S5P but not S2P at active PRC-targets. Ring1B, S5P and S2P occupancy at Active (*ß-actin*), Inactive (*Myf5*) and PRCa genes were analysed <1kb from TSSs by ChIP or re-ChIP and qPCR, as described in (Figure 5.14). Background levels (mean enrichment after first ChIP with Ring1B or re-ChIP with no antibody in the second IP) shown as white bars. Mean and standard deviations presented from 2-5 independent biological replicates. ChIP demonstrates binding of Ring1B, S5P and S2P to PRCa genes, but re-ChIP shows PRC1 co-association with S5P but not with S2P above background levels (white bars). Experiments were performed by E. Brookes.

Interdependence of RNAPII and PRCs in the PRC-repressed state is further shown by a strong positive correlation between S5P and either H2Aub1 or H3K27me3 promoter levels within the PRCr group (Spearman's correlation coefficient = 0.64 and 0.52,

respectively; Figure 5.22) which is consistent with interplay between RNAPII S5P and PRCs in the PRC-repressed chromatin. Interestingly, at PRCa genes, similarly to Active genes, S5P shows lower correlation with H2Aub1 and H3K27me3 (Spearman's correlation coefficient = 0.04 and 0.02, respectively), but increased correlation with S2P (Spearman's correlation coefficient = 0.28; Figure 5.22), consistent with its presence with active RNAPII (S2P). The low correlation between S5P and H2Aub1/H3K27me3 at PRCa genes is consistent with Model 1 (Figure 5.1) and the existence of a mixture of states across the cell population, one state characterised by PRC-repression where S5P and PRC1 and PRC2 co-associate, and another state characterised by the co-occurrence of S5P and S2P, in which the gene is actively transcribed. The PRCi group displays an intermediate correlation between S5P and H2Aub1 or H3K27me3, which may be due to a lower proportion of active alleles within the population, allowing the PRC-bound state to contribute more to the correlation.

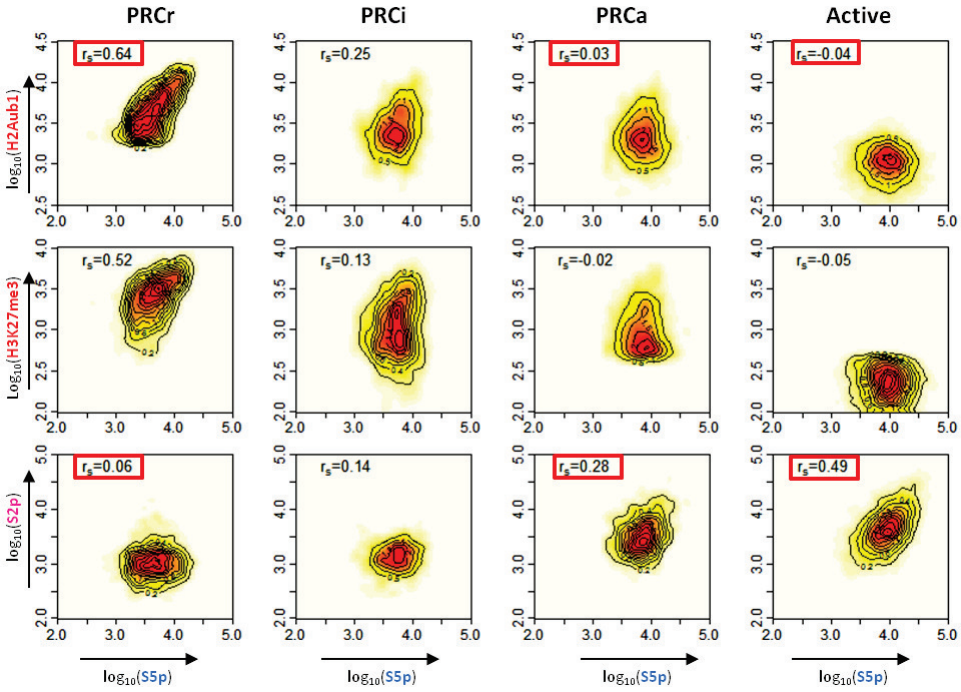


Figure 5.22: Strong positive correlation between S5P and H2Aub1/H3K27me3 within PRCr and between S5P and S2P at Active within PRCa genes. Sequencing depth was calculated in 2kb windows centered at genes TSS (H2Aub1, H3K27me3 and S5P), or downstream of TTS (S2P). Scatter plots are displayed in the log10 scale for visualisation purposes. To avoid taking the log of 0, the value 1 was added prior taking the log. Spearman's correlation coefficient is depicted in each plot (r_s).

5.3.15 *Pluripotent transcription factors are enriched at PRCa genes*

The stem cell-specific transcription factors Oct4 and Nanog direct ES cell identity by regulating expression of pluripotency- or differentiation-specific genes. A subset of genes regulated by the core transcriptional network is known to be co-occupied by PcG proteins (Loh et al., 2006, Boyer et al., 2005, Lee et al., 2006). To explore whether pluripotency factors might influence the switch between PRC-repression and activity, I analysed two published data sets of genome-wide binding sites of transcription factors associated to pluripotency (Kim et al., 2008, Chen et al., 2008).

Chen et al. [2008] used ChIP-Seq and commercially available antibodies to identify the binding sites of several proteins, including including "core" pluripotency factors (Nanog, Oct4, and Sox2; Loh et al., 2006, Boyer et al., 2005), somatic cell reprogramming factors (Klf4 and c-Myc; Takahashi and Yamanaka, 2006), known to regulate the expression of pluripotency genes (Esrrb, Zfx, n-Myc, Tcfcp2l1); E2f1, a regulator of cell cycle progression; and two other proteins necessary to maintain cultured ES cells in the pluripotent state (BMP-induced Smad1 and LIF-induced Stat3). Typically, ChIP-Seq data analysis of transcription factor binding sites involves the definition of a cuttoff maximum distance to assign a particular binding site (peak) to the nearest transcription initiation site. To avoid the use of an arbitrary threshold, Chen et al. [2008] calculated an association score which defined the probability of a given TFBS to be associated to its proposed target genes. We used Chen et al. [2008] datasets to test for enrichment of TF target genes in each of the 6 clusters of genes (Active, PRC and Inactive clusters). To do so, I performed GSEA (Subramanian et al., 2005) using a total of 13 ranked gene lists, one for each TF (Nanog, Oct4, Sox2, Klf4, c-Myc, Esrrb, Zfx, n-Myc, Tcfcp2l1, E2f1, Smad1, STAT3 and CTCF). Given a ranked list of TF scores, GSEA tests whether the members of each cluster are randomly distributed along the list or particularly found near the edges, with the latter indicating enrichment for targets of the TF in question. An example of a GSEA plot is shown in Figure 5.23 where it is seen that Active and PRCa genes are significantly associated with high Nanog scores (FDR < 25%), indicating that genes within the Active and PRCa groups are enriched for Nanog targets. This analysis demonstrated that all analysed TFs are significantly enriched within the Active group (Table 5.3).

As for the PRC-cohorts, PRCa group was enriched for many of the analysed TFs (Nanog, Sox2, Klf4, Esrrb, Tcfcp2l1, Smad1, Stat3 and CTCF) suggesting the coordinate involvement of these TFs and Polycomb complexes to maintain the pluripotent state and repress lineage-specific genes. Previous studies have already shown that Nanog, Sox2, and Oct4 share many common targets with PRCs (Boyer et al., 2005, Lee et al., 2006). PRCi cluster is particularly enriched for Klf4, Esrrb and Tcfcp2l1, while PRCr genes are particularly associated with CTCF and Esrrb binding cites. The constitutive transcription factor CTCF has been previously linked to Polycomb function (Li et al., 2008). Interestingly E2F1 binding, a known cell cycle regulator (Iaquinta and Lees, 2007) whose expression level varies across the cell cycle, is enriched within the PRCa group. Binding motifs of E2F1 have already been shown to be particularly enriched in PRC-targets (Prediction of Polycomb target genes in mouse embryonic stem cells). The role of Polycomb complexes in the regulation of cell cycle, if any, is yet to be explored.

Table 5.3: Gene set enrichment analysis (GSEA) of TF targets in six clusters of genes. FDR is represented and highlighted if a gene group is significantly enriched for targets of a particular TFs (FDR < 25%).

	FDR (significant if FDR < 0.25)					
TFs	Inactive	PRCo	PRCr	PRCi	PRCa	Active
Nanog	1	1	1	1	0.056	0
Oct4	1	1	1	1	0.302	0
Sox2	1	1	1	1	0.021	0
Klf4	1	1	1	0.050	0.215	0
c-Myc	1	1	1	1	1	0
Esrrb	1	1	0.160	0.025	0	0
Zfx	1	1	1	1	0.737	0
n-Myc	1	1	1	1	0.754	0
Tcfcp2l1	1	1	1	0.075	0.017	0
E2F1	1	1	1	1	0.002	0
Smad1	1	1	1	1	0.015	0
Stat3	1	1	1	1	0.018	0
CTCF	1	1	0.138	0.01	0.026	0

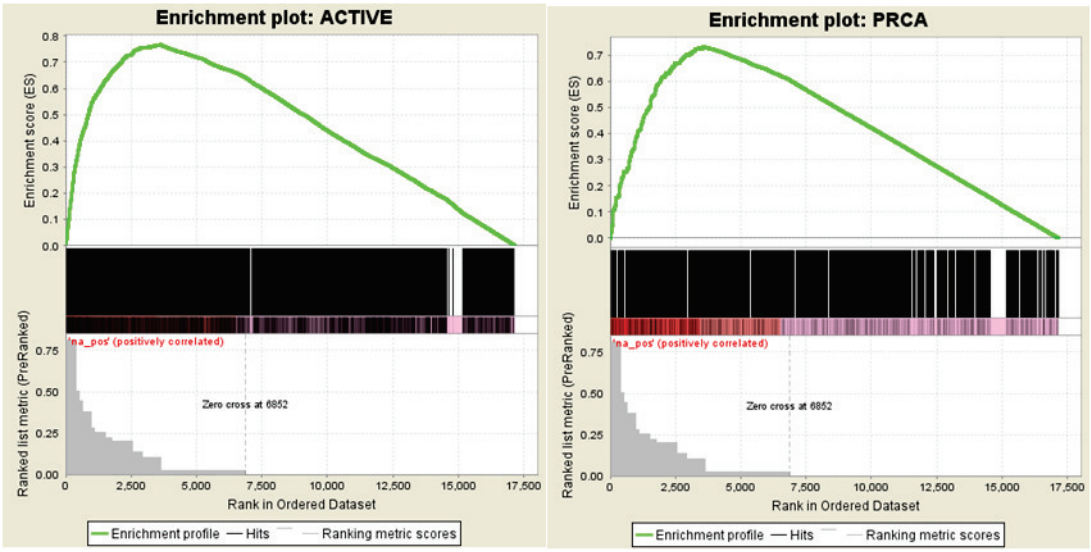


Figure 5.23: GSEA plot showing enrichment for Nanog targets in PRCa and Active groups. Significant enrichment of Nanog targets in the PRCa and Active groups of genes. The ranked gene list of Nanog scores is shown in the lower, grey plot. Vertical black lines indicate where genes in the Active or PRCa groups fall in the ranked gene list. The top, green line indicates the running enrichment score (ES) that becomes more positive as genes from the top list match the ones in the ranked list. Enrichment for Nanog targets was not significant in the PRCi/r/o and inactive groups of genes (i.e. FDR < 25%).

Kim et al. [2008] have applied *in vivo* biotinylation mediated ChIP (bioChIP) to identify target promoters of nine transcription factors associated with pluripotency of ES cells (J1 mouse ES cells), including some of the proteins previously examined by Chen et al. [2008] (Oct4, Sox2, Nanog, Klf4 and c-Myc), transcriptional regulators known to interact with Nanog (Nrb1, Nac1 and Zfp281; Wang et al., 2006), and the Oct4 target Zfp42 (also known as Rex1). This approach relies on streptavidin affinity capture of tagged proteins thus avoiding issues related to antibody availability. The authors report that target genes fall into two classes: promoters bound by few factors (less then four), which tend to be inactive or repressed in the pluripotent state; and promoters bound by four or more than four factors which are largely active in the pluripotent state and become repressed upon differentiation (Kim et al., 2008). Therefore, we focus our analysis of Kim et al. [2008] datasets on the examination of targets of six TFs that exhibit overall similarity in their targets (Nanog, Sox2, Dax1, Nac1, Oct4 and Klf4; Zfp281 was excluded due to less target gene overlap). We verify that genes with multiple factors are enriched for active and PRCa genes, while gene targets of a single TF are enriched for targets of Inactive, PRCo and PRCr genes

(Figure 5.24). Among all gene clusters, more then half (54% - 77%) of the genes are often bound by 1 TF. PRCa and Active genes (15% and 13%, respectively) are more often bound by $\geq$4 TFs then PRCi, PRCr, PRCo and Inactive genes (9%, 9%, 5%, 10%, respectively).

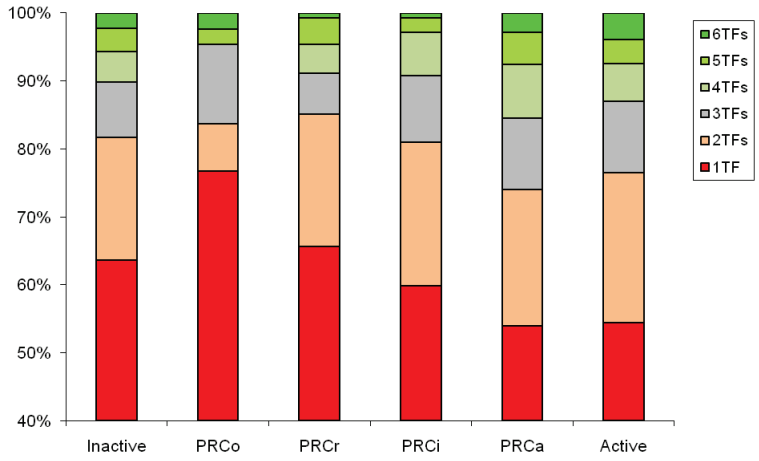


Figure 5.24: Distribution of multiple/single TFs target genes among Active, PRC cohorts and Inactive groups of genes. Percentages against the total number of genes (e.g. "Nr of Active genes with 6TFs" / "total nr of genes with 6TFs") is shown in the Y-axis.

5.3.16 *Single cell analysis of PRCa gene, Lefty2, by immuno-cryoFISH*

To investigate whether the association of PRCa genes with PRCs or with active (S2P) RNAPII was dependent on Nanog levels across the ES population, we performed single-cell co-localisation experiments of the PRCa gene *Lefty2*, with RNAPII-S5P, RNAPII-S2P and Ezh2 (PRC2) in ES cells co-stained with Nanog antibodies (Figure 5.25), using cryoFISH. The experiments confirmed the association of *Lefty2* gene with RNAPII-S5P independently of the Nanog levels, but interestingly, there was a statistically significant preference for the association of *Lefty2* locus with RNAPII-S2P in Nanog^{high} cells and with Ezh2 in Nanog^{low} cells (p-value=0.1 (S5P), 0.003 (S2P), and 0.02 (Ezh2) for comparisons between Nanog^{high} and Nanog^{low}, Chi^2 test). This suggests that Nanog, and possibly other pluripotency transcription factors that fluctuate within the ES cell population, might regulate the switch between PRC-repressed and active states of PRCa genes.

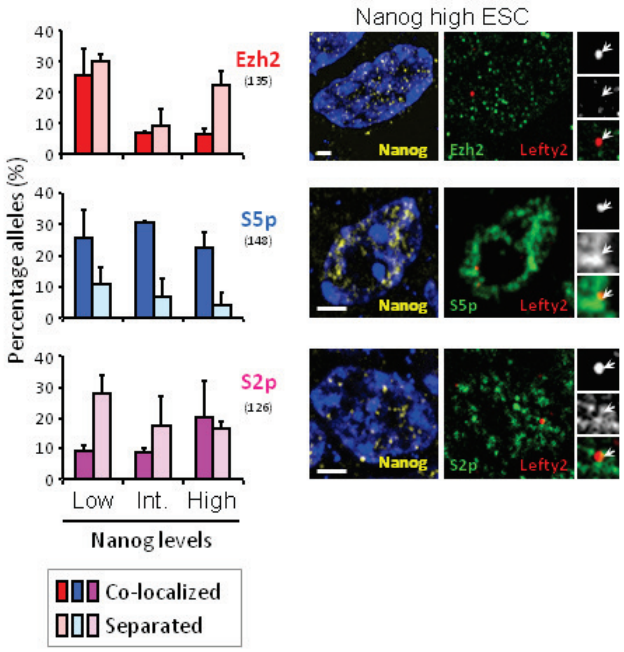


Figure 5.25: PRCs and elongating RNAPII-S2P are mutually exclusive at the *Lefty2* gene. The frequency of association of the PRCa gene *Lefty2* (pseudo-colored red, arrows) with sites containing S5P, S2P or Ezh2 (green) was measured in ESC nuclei expressing different levels of Nanog (yellow; classified as 'low', 'intermediate' or 'high') by immuno-cryoFISH. Locus association with each marker was scored as 'co-localised' ($\geq$1 pixel overlap between red and green signals) or 'separate'. The *Lefty2* locus associates with S5P at similar frequency regardless of Nanog levels, but association with S2P is highest in Nanog^{high} with and Ezh2 is highest in Nanog^{low} cells. Bar: 2 µm. Number of loci analysed indicated in brackets. Experiments performed by S. Xie.

5.4 DISCUSSION

In this chapter, molecular, cellular and genomic analyses on a single gene and genome-wide scale were utilised to explore the relationship between PRC and RNAPII recruitment and gene expression. We have found that a proportion of PRC-targets are occupied by unproductive RNAPII complexes but a proportion of them also exhibit full-length transcription and the concomitant presence of RNAPII-S2P and H3K36me3, two hallmarks of transcription elongation. This study suggests that while PRC repression is associated with a single RNAPII state (S5P+S7P-S2P-) across all CpG-rich genes, fluctuations between the PRC-repressed and the canonical active state (S5P+S7P+S2P+) occurs to variable extents, resulting in differing expression levels and in the identification of "active" PRC-repressed genes.

PRC targets whose state switches between active and repressed across the cell population, potentially in response to pluripotency transcription factors, are involved in metabolic processes and signaling pathways that are important for ES cell biology and likely crucial for specific and efficient lineage commitment. It will be important to further explore the regulation of PRC-repressed RNAPII complexes, in order to understand and manipulate pluripotency.

5.4.1 *The three major epigenetic states in ES cells: Active, PRC-repressed and Inactive*

Developing a binary code for RNAPII, H3K27me3 and H2Aub1 histone marks (Chapter 4) allowed us to investigate the combinatorial association of each gene with RNAPII modifications or PRC and the identification of six different groups of genes by hierarchical clustering (Figure 5.2): Active genes; Inactive genes, and four different groups of PRC-regulated genes. Four distinct cohorts of PRC-target genes were characterised by the investigation of RNAPII modifications, epigenetic and expression states (Figure 5.7). Within the different PRC cohorts we identified: PRC-only (PRCo) genes with H3K27me3 and no recruitment of RNAPII; PRC-repressed (PRCr) genes with PRC and RNAPII S5P; PRC-intermediate (PRCi) genes with PRC, S5P, S7P and 8WG16; and PRC-active (PRCa) with PRC, S5P, S7P, 8WG16 and S2P. In this study we hypothesize that ES cell genome is characterised by three distinct states of gene regulation: Active (CpG-rich promoters, high RNAPII and high expression); PRC-repressed (CpG-rich promoters, high RNAPII and low expression); and Inactive (CpG-poor promoters and low or absence of RNAPII and expression).

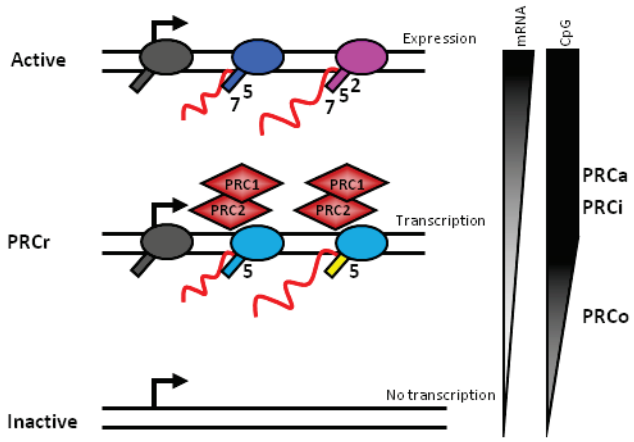


Figure 5.26: Model of gene regulatory states in ES cells. Three states are identified: active, PRC-repressed and inactive. Active genes are associated with S5P (blue) during initiation, S2P (pink) during elongation, productive expression and no PRC-association. PRC-repressed genes recruit RNAPII-S5P, which extends into coding regions without S2P, which may require additional modifications (yellow). PRCr genes are transcribed but mature mRNAs are extremely low or not detected. Inactive genes are not associated with RNAPII or PRC. Between the active, PRC-repressed and inactive states, increasing activity is correlates with associated epigenetic signatures and an increase in all RNAPII forms. PRCo, PRCi and PRCa gene classes occupy intermediate levels between the continuum behavior from the active to the inactive spectrum of the scale.

5.4.1.1 *Active state*

The group of Active genes comprised a total of 5707 genes which show characteristics of the previously described 20% most expressed genes (Top20%; n = 3772 genes), with high levels of S5P, 8WG16, S7P and H3K4me3 at their promoters, and high levels of elongation marks S2P and H3K36me3 throughout gene bodies (Figure 5.7). In fact, 90% of Top20% genes (2306 out of 2564) are inside this group. The group of Active genes is associated with housekeeping and general functions of the cell such as “ribosome”, “DNA repair” and “nucleic acid metabolic processes”. Promoters belonging to this group have characteristics of previously identified characteristics of housekeeping genes (high expression, CGI promoters, broad promoters; Carninci et al., 2006) .

5.4.1.2 *PRC-repressed genes*

We have identified three groups of PRC-regulated genes (1632 PRCr, 742 PRCi, and 1227 PRCa genes) at which there is RNAPII recruitment and a group of PRC genes (denominated PRCo, 798 genes) that do not recruit RNAPII. In particular, the initiating form of RNAPII (S5P) is recruited to the promoters of all genes belonging to all three classes of genes that recruit RNAPII (PRCr,i,a;). Similarly, CpG islands are over-represented at the three groups of genes but not at the PRCo group (Figure 5.9), and a proportional increase of mRNA expression is detected from PRCo to PRCa genes (Figure 5.5). Accompanying mRNA expression, an increase of levels of RNAPII modifications associated with gene activity (8WG16, S7P and S2P) is also observed from the PRCi to PRCa groups (Figure 5.19). The group of PRCr genes shows characteristics similar to the majority of the previously described bivalent genes with no S7P, 8WG16 or S2P detection, while RNAPII-S5P is detected at the promoters and throughout the coding region of PRCr genes, but no increase in RNAPII after the 3' end, unlike at Active genes. Absence of S2P and S7P from PRC-repressed RNAPII, together with the lack of RNAPII beyond TTSs, indicates that PRC-repression may interfere with RNA processing and the production of mature mRNAs.

5.4.1.3 *Inactive genes*

A cohort of 5298 Inactive genes was identified. These genes showed no recruitment of RNAPII or active histone marks (H3K4me3, H3K36me3) or PRC proteins or PRC-mediated modifications (H3K27me3, H2Aub1). Gene ontology analysis showed an enrichment for terms associated to "multi-cellular organismal process" and "signal transduction" suggesting that the activation of these genes is most relevant in somatic tissues. This group comprised a total of 5298 genes which corresponds to 28% of the genome, demonstrating that in ES cells a percentage of the genome is maintained in the fully repressed state, possibly by mechanisms such as DNA methylation, or the existence of other histone modifications such as H3K9me2 or H3K20me3 which are known to be involved in mechanisms of chromatin compaction (Lu et al., 2008). DNA methylated promoters were statistically enriched at the group of Inactive genes (p-value = 4.6×10^{-15}; Figure 5.9), which is also populated by promoters of low CpG

content (Figure 5.9). Low CpG content at this genes may explain their lack of RNAPII recruitment (Rozenberg et al., 2008) and their sharp promoter type which is commonly seen at CpG-poor promoters (Carninci et al., 2006). These features are characteristic of tissue-specific genes and suggest that specific transcription factors are required to initiate transcription at the appropriate differentiated cells.

5.4.2 *CpG promoter structure and the regulation of gene expression*

CpG promoters constitute the majority of the mammalian genes (approximately 60% of the promoters considered in this study overlap CpG islands). We have demonstrated that gene classes which recruit RNAPII are also over-represented by CpG island promoters (CGI promoters; Active, PRCa, PRCi, PRCr; Figure 5.9) and, by the use of CAGE datasets, we have shown that these genes display broad type of promoters (Figure 5.11) in which initiation can occur at a number of sites. Indeed, it has been suggested in the literature that CGI promoters adopt a transcriptionally permissive state and display dispersed initiation patterns (Carninci et al., 2006, Juven-Gershon et al., 2008). Accordingly, several studies have revealed that CGI chromatin is prone to transcriptional initiation (reviewed in Deaton and Bird, 2011), including nuclear run-on experiments that have detected bidirectional transcription of short unproductive RNAs at many CGI promoters (Core et al., 2008, Seila et al., 2008). Other studies have detected transcriptional activity even at intergenic CpG islands or CpG islands located outside transcriptional units. Examples of this phenomena are the ncRNAs *Air* and *Kcnq1ot1*, which initiate at a CGI within an intron of *Igf2r* and *Kcnq1* genes, respectively (Sleutels et al., 2002, Mancini-DiNardo et al., 2003).

The classes of genes with the highest CpG content promoters also showed the lowest density of histone H3 (Figure 5.10) which suggests that these genes are intrinsically accessible and less reliant of ATP-dependent chromatin remodelers like the SWI/SNF complexes. An open chromatin configuration at the CpG-rich promoters may be either favorable for, or a consequence of, RNAPII recruitment and transcription initiation from broad promoters, or both. Indeed, it has been suggested that the association of the transcription machinery plays a decisive role in determining the positioning and occupancy of nucleosome (Schones et al., 2008, Gilchrist et al., 2008),

preventing nucleosome formation and facilitating further gene activation (Gilchrist et al., 2010).

In support to our findings, CpG-rich sequences have been related to the recruitment of RNAPII (Rozenberg et al., 2008). Further support comes from a study from Thomson et al., 2010 which have shown that Cfp1 protein (a component of the Setd1 H3K4 HMT complex) selectively binds to non-methylated CGI sequences *in vitro*. It is known that RNAPII-S5P can also recruit H3K4 HMTs directly to chromatin (Ng et al., 2003), which in turn creates a "open" chromatin environment that facilitates the binding of RNAPII and necessary transcription factors (Mohn and Schubeler, 2009), thus it is likely that a cascade of positive feedback loops reinforces the prevalent epigenetic state of CGI promoters.

Paradoxically, CpG sequences also play a role in the recruitment of PRC2 (Tanay et al., 2007, Mendenhall et al., 2010). Indeed, three of the described PRC classes (PRCr, PRCi and PRCa) are bivalent (H3K4me3 and H3K27me3) and CpG-rich. However, on the contrary to the majority of PRC targets, PRCo genes do not possess CpG islands and have low level of GC content or CpG frequency. Thus, the existence of PRCo class of promoters suggests that these genes recruit PRC2 in a manner that is not dependent on the CpG content of the promoter. PRCo promoters also do not recruit RNAPII and lack H3K4me3 mark supporting the view that RNAPII and H3K4me3 are mechanistically linked (Ng et al., 2003).

5.4.3 *The merging of Polycomb- and DNA methylation-mediated gene silencing at PRCo genes*

DNA methylated promoters were over-represented within the groups of Inactive and PRCo genes. Interestingly, a reasonable proportion of PRCo genes (37%; Figure 5.9) are methylated, and thus may be silent by two epigenetic mechanisms, DNA methylation and Polycomb repression. Although differences in ES cell lines between the cross-referenced studies (ES-V6.5 and ES-J1, used by Mikkelsen et al., 2007 and Fouse et al., 2008, respectively) have to be taken into account, the observation that DNA methylation and PcG silencing can be coupled has been shown previously by Vire et al. [2006] where it was shown that Ezh2 Polycomb group protein can recruit

DNA methyltransferases directly to certain genes both *in vivo* and *in vitro* in HeLa cells (Vire et al., 2006). In support to the mechanistic link between Ezh2 and DNA methylation, it has been found that H3K27me3 marked promoters in ES cells are more likely to acquire DNA methylation during development (Mohn et al., 2008) and in cancer (Ohm et al., 2007, Widschwendter et al., 2006). The functional relevancy of DNA methylated marks at PRC2-targets is not known; perhaps DNA methylation may help to "lock in" the silent state by adding another layer of repression, or might even be relevant in the inheritance of Polycomb chromatin during DNA replication (Taghavi and van Lohuizen, 2006).

5.4.4 *Promoter-proximal pausing versus "poised" RNAPII at PRC-repressed genes*

Promoter-proximal pausing of RNAPII was originally described at silent stimulus-response heat-shock genes (Rougvie and Lis, 1988), which are thought to pre-load RNAPII to allow quick response to appropriate signals. Pausing occurs after RNAPII transcribes approximately 25 to 50 nucleotides and its facilitated by the DSIF and NELF complexes (reviewed in Levine, 2011). Pausing can be clearly advantageous for quickly inducible genes, however, this rate limiting step of transcription is not only relevant at stimulus-response genes, with even highly transcribed genes showing a significant degree of paused RNAPII in mammals (Core et al., 2008, Rahl et al., 2010, Min et al., 2011) and in *Drosophila* (Gilchrist et al., 2008, 2010). Gilchrist et al., 2010 suggested that pausing serves as an early "checkpoint" before the release of RNAPII into productive elongation, at some genes (e.g. active genes) the pausing state can be transient, while at others (e.g. heat-shock genes), it may involve the formation of a paused complex that represses gene expression until the appropriate signal is received.

The pre-loading of RNAPII at bivalent genes in ES cells is thought to provide the flexibility needed to prime these genes for activation or repression upon lineage specification (Bernstein et al., 2006, Mikkelsen et al., 2007). However, while RNAPII pausing is also classified as a mechanism of gene priming, the characteristics of paused RNAPII complexes are different from the RNAPII complexes found at PRC-repressed genes. For instance, at paused genes low levels of RNAPII are

observed in the promoter-proximal region, while at PRC-target genes, considerable amounts of RNAPII-S5P are detected at promoters and coding regions. Furthermore, contrasting to what we observe at PRCr genes, RNAPII is recognised by 8WG16 at paused genes (Kim et al., 2005, Guenther et al., 2007, Boehm et al., 2003).

A clear group of "paused" genes was not detected in our analysis. This is in apparent conflict with an earlier report from Guenther et al. [2007] in which the authors detect approximately 35% of silent genes with RNAPII (detected by 8WG16 antibody) enriched at the 5′ end but with no evidences of transcript accumulation, in human ES cells. This observation led the authors to propose transcriptional pausing, poor processivity, or abortive initiation at these genes. The difference between our study and Guenther et al. [2007] can be explained by the sensitivity in detecting transcripts by mRNA-Seq compared to the microarray expression assay used by the authors. For instance, low levels of transcription would be scored as absent leading the authors to conclude that these genes are silent, nonetheless experience transcription initiation.

It is possible that the paused state detected in *Drosophila* is not a prevalent mechanism of priming in ES cells in mammals. Similarly, the characterised PRC-repressed state, with RNAPII extending through the coding regions of PRC-repressed genes, may not exist in *Drosophila*. For instance a study in *Drosophila* embryos identified RNAPII-S5P and -8WG16 at the promoters of PRC-repressed *Hox* genes, but not throughout the gene bodies (Chopra et al., 2009b). Interestingly, bivalency has not been observed in lower organisms such as *Xenopus* and *Drosophila*, raising the possibility that the poised state of RNAPII at PRC-bound genes evolved recently and may be specific to mammals.

5.4.5 *Synergy between PRCs and RNAPII at PRCr genes*

In the past, much emphasis has been given to the recruitment of RNAPII to promoters as a key regulatory step of gene expression (Nevado et al., 1999). More recently, a plethora of studies has revealed that a large number of genes are regulated at steps downstream of the recruitment of RNAPII to a promoter. Namely, several genome-wide studies in human (Guenther et al., 2007, Kim et al., 2005, Rahl et al.,

2010) and *Drosophila* (Muse et al., 2007, Zeitlinger et al., 2007) discovered significant proportions of inactive genes with high levels of RNAPII at their promoters.

At PRCr genes, remarkably similar S5P and PRC distributions at TSSs and throughout coding regions (Figure 5.12, Figure 5.13) suggest that PRCs travel through genes coincidental with elongating S5P complexes. RNA from PRCr genes was detected, at low levels (Stock et al., 2007), and few transcripts are made per cell (Guenther et al., 2007). In addition, despite the presence of RNAPII-S5P through the gene bodies of many PRCr genes, absence of RNAPII-S2P and H3K36me3 suggest that efficient processing of transcripts is limited. Therefore, the deposition of H2Aub1 and/or H3K27me3 marks may act to block proper efficient transcription at some level. This is consistent with the role of Ring1B (Stock et al., 2007) and H2Aub1 (Zhou et al., 2008) in negative transcriptional regulation. Stock et al. [2007] showed that conditional depletion of Ring1B results in gene de-repression, while Zhou et al. [2008] showed that the H2Aub1 at promoters prevents the recruitment of FACT resulting in the inhibition of transcriptional elongation. Nonetheless, PRC-regulated genes are not completely silent. We and others have found low level transcription from bivalent genes (Stock et al., 2007, Core et al., 2008, Min et al., 2011). Therefore, another possibility that may account for the synergy between PRCs and RNAPII is that RNAPII transcription products may have direct roles in the recruitment or stability of PRCs. For instance short ncRNAs transcribed from 5' end of PRC2-targets in both ES and T cells which have been shown to interact with PRC2 and to cause gene repression in *cis* (Kanhere et al., 2010).

Only a detailed analysis of transcript products of PRC-targets genes, including splicing patterns and orientation of transcription, would help to resolve this open question.

5.4.6 *Active PRC-repressed genes (PRCa, PRCi) as a read-out of heterogeneous ES cell populations*

Heterogeneity of stem cell populations is a phenomena characterised by the existence of a variety of inter-convertible states within the ES cell population, under pluripotent conditions, that are responsive to fluctuations in the transcription factor network.

These phenomena have been recently addressed (Graf and Stadtfeld, 2008, Canham et al., 2010). Several transcription factors are known to exhibit such fluctuations in expression across the undifferentiated ES cell population, such as *Nanog*, *Zfp42* (*Rex1*) and *Stella* (*Dppa3*) (Chambers et al., 2007, Hayashi et al., 2008, Toyooka et al., 2008).

In this chapter we suggest that a proportion of the characterised PRC-target genes exists in different states across the ES cell population. Several lines of evidence support the idea that the "active" PRC-repressed state represented by PRCi and PRCa genes reflect a mixture of states. First, sequential ChIP experiments demonstrated that although Ring1B, Ezh2 and S5P co-localise at PRCr, PRCi and PRCa genes, Ring1B and S2P did not co-localise (9/9 genes), suggesting that they do not share the same chromatin fragment (Figure 5.21). Thus, the concomitant detection of S2P and H2Aub1 or H3K27me3 by ChIP-Seq experiments could simply reflect an artifact of this type of sampling experiments, i.e, ChIP-Seq detects the average distribution of factors at a given gene across a population of cells/alleles and could therefore mask what could be different, separated, states that exist either in different cells or even in different alleles of the same cell. Secondly, at PRCr genes H2Aub1 and H3K27me3 marks correlate positively with RNAPII-S5P, while at PRCa genes (similarly to active genes) the correlation between S5P and H2Aub1 or H3K27me3 is partially lost while the correlation with S2P increases, thus supporting the idea of two separate S5P states and that the elongating form of RNAPII (S5P+S2P+) does not co-associate with Polycomb complexes (Figure 5.22).

It has been suggested that fluctuations in levels of factors of which expression levels fluctuate across ES cell population may prone ES cells into a specific lineage of differentiation, however this phenomena is poorly characterised (Graf and Stadtfeld, 2008, Canham et al., 2010). Nanog is a known regulator of ES cell pluripotency and its expression levels fluctuate within the ES cell population, resulting in a heterogeneous pattern of expression (Chambers et al., 2007). Interestingly, Nanog target genes are enriched at PRCa class (Figure 5.23). Furthermore, CryoFISH experiments of a PRCa gene, *Lefty2*, shows that *Lefty2* associates with RNAPII-S2P but not Ezh2 in Nanog^{high} cells, and the opposite, associating with Ezh2 but not RNAPII-S2P, in Nanog^{low} (Figure 5.25). This suggests that Nanog, possibly working in concert with other

pluripotency transcription factors that fluctuate within the ES cell population, might regulate the switch between the PRC-repressed and active states of PRCa genes.

Taken together, our data suggests that the "mixed" ChIP profile detected for the PRCi and PRCa genes reflects a mixture of relative proportion of cells in which a gene is either active (with RNAPII S5P+S2P+ producing full-length mRNA that leads to protein expression), or PRC-repressed (with RNAPII S5P+S2P- co-associated with Polycomb). The regulation of these genes might be important for developmental plasticity. Understanding the dynamic fluctuations in levels of important transcription factors will allow us to better understand how gene regulation is achieved in the pluripotent state, which will definitely contribute to our ability to manipulate the conversion between pluripotent and differentiated states, and vice-versa.

5.4.7 *Interactions between PRCs and the transcriptional network*

The observation of shared targets between PRCs and pluripotent TFs, namely Oct4, Sox2 and Nanog, has been previously reported in human (Lee et al., 2006, Boyer et al., 2006) and mouse (Squazzo et al., 2006) ES cells, where these TFs have been indicated to act both as a repressors or activators (Boyer et al., 2005, Loh et al., 2006). Mechanisms that account for the differential regulation of TFs in ES cells are not well understood. It was suggested that the differential regulatory effects of TFs result from differences in the combinatorial binding of several TFs (Kim et al., 2008). For instance, Nanog may bind to targets that are repressed by Polycomb and act to repress its targets, or it may act cooperatively with other TFs to trigger gene activation. A more recent study from Ouyang et al., 2009 goes further in the identification of a specific set of TFs that are more often associated with PRC2-target genes, namely Oct4, Nanog, Sox2, Smad1, Stat3, Tcfcp2l1, and Esrrb. The authors identify a simple rule for the mode of cooperation among different TFs which is explained by the existence of two groups of TFs: group I, which includes E2F1, c-Myc, n-Myc and Zfx and is associated to the activation of its target genes; and group II TFs which includes Oct4, Nanog, Sox2, Smad1, Stat3, Tcfcp2l1, and Esrrb and can have either activating or repressive roles, i.e., contributing to the repression of some genes, but if cooperating with group I TFs contribute to the activation of their target genes.

An important caveat of this study is that the classification of "PRC2-targets" used by Ouyang et al., 2009 relies solely on Suz12 binding data published in Chen et al., 2008. Closely inspection of this dataset reveals that only a small proportion of PRC2-targets is represented in this set, and the majority correspond to PRCr genes. Therefore, the conclusions of Ouyang et al., 2009 could apply to the mode of regulation operating in the PRCr group of genes but cannot be extended to PRCo, PRCi or PRCa groups.

Our data reveals that distinct PRC-target genes can be bound to either fewer (e.g. PRCr) or multiple (e.g. PRCa) TFs (Figure 5.24). Thus, one could speculate that at the class of PRCr genes, TFs are weak activators and could even function in association with PRCs to repress its targets (as suggested by Wang et al., 2006), while at PRCa genes the binding of multiple TFs would result in gene activation and the gene would switch from a PRC-repressed state to an active state. To provide an independent validation of this model, we are currently planning experiments using Nanog-/- cells. According to this model, knockdown of Nanog expression would down-regulate the expression levels of genes in which it acts as an activator (e.g. PRCa and Active genes) and up-regulate the expression levels of genes in which it acts as a repressor (e.g. PRCr genes).

Although we have not tested the co-localisation of Nanog, Oct4 or other TFs with subunits of the PcG group of proteins at PRCr genes (as these would likely be mediated by long range chromatin interactions), the localisation of PRCs at targets of the core transcription factors (Nanog, Oct4 and So*x*2) detected by ChIP-Seq indicates that the Polycomb and pluripotency transcription factors may cooperate in the repression of differentiation specific genes. A possible link between core factors and PRCs was revealed by Wang et al. [2006] who co-purified Oct4, YY1 and Ring1b with Rex1 from mouse ES cells. It would be interesting to investigate the epigenetic state of ES cell populations using Nanog expression reporters and conditional knockout of TFs.

6

SEQGI - A SOFTWARE TO ANALYSE CHIP-SEQ DATA

6.1 RESEARCH MOTIVATION

Analysis of high-throughput sequencing data often requires the visualisation and quantification of read density and read patterns over classes of genomic features of interest as well as the statistical testing of differential enrichment between classes of genomic features. Furthermore, the increasing use of multiple biological samples (e.g. different tissues, knockouts, differentiated *versus* undifferentiated conditions, etc.) demand for statistical and computational methods suitable for the assessment of biological variation. The processes of visualisation and statistical testing have largely remained disparate and so have led to the use of multiple softwares requiring inter-conversion between differing file formats.

In this project of my PhD, I aimed to develop a open-source application that provides users with an intuitive graphical interface, combining both visualisation and statistical tools, and thus that can be used to assist in the rapid interpretation of sequencing data.

6.2 STATE OF THE ART

6.2.1 *Quantifying ChIP-Seq signal ("counts") versus identifying peaks*

ChIP-Seq is a powerful method used to map epigenetic modifications or binding sites of DNA-associated proteins throughout the genome at an unprecedented resolution. A typical ChIP-Seq run generates gigabases of sequences that, once properly aligned to a reference genome, essentially represent a continuous signal of enrichment for a given analysed factor.

The type of ChIP-Seq enrichment can very between different types of proteins. For instance, the insulator binding protein CTCF exhibits sharp peaks at its binding sites (Park, 2009), while profiles for RNAPII and histone marks such as H3K27me3 are characterised by broader regions of enrichment. Typical data analyses confides on peak-finder softwares that convert the continuous signal into discrete events of enriched/non-enriched regions to allow the localisation of specific binding or marked sites across the genome. However, most software available is optimised for the detection of peak-type data and currently there are few available programs or packages, such as CCAT (Xu et al., 2010) and SICER (Zang et al., 2009), which take into account broader domains of enrichment. Moreover, peak-finders disregard the continuous values of enrichment which can provide an important, and sometimes crucial, level of information to the investigation. For instance, continuous values can allow the study of progressive changes of a given epigenetic mark between different conditions, similarly to the type of analysis typically performed with mRNA-Seq data in which temporal changes in gene expression are studied across conditions. This, however poses another challenge in ChIP-Seq data analysis which is the need to normalise the datasets before performing such comparisons. The issue of data normalisation and differential expression has been extensively debated in the context of mRNA-Seq data analysis (Bullard et al., 2010, Oshlack et al., 2010), and it is thought that some of the developed methodologies can also be applied to ChIP-Seq count-type data (Anders, 2010). Other examples of the value of using read counts to quantify marker enrichment, instead of discreet present/absent classifications, come from our own investigation. For example, by comparing H2Aub1 and RNAPII-S5P promoter levels, we found a strong positive correlation between the two modifications at PRCr but not PRCa genes, which, by complementing with a experimental validation, proved to represent a biologically meaningful association.

6.2.2 *Current available packages to analyse and visualise ChIP-Seq count data*

One of the greatest challenges in the analysis of high-throughput sequencing datasets is the amount of data that is generated in each single run, and even after the first processing steps (e.g. after genome alignment), advanced bioinformatic skills are still required to handle the data generated. Efforts to address this challenge have been

made by the bioinformatic community, but there are still few approaches available. Galaxy (Giardine et al., 2005) is an example of a bioinformatic tool that can be easily used to calculate read counts at a set of genomic intervals (e.g. promoters). However, a downside of Galaxy is that it depends on the loading of the datasets into a common database, which could be time limiting for large datasets such as those derived from high-throughput sequencing experiments. Other stand-alone, downloadable, applications such as BEDtools (Quinlan and Hall, 2010) and SAMtools (Li et al., 2009) overcome this problem, however both depend on the use of command line and are not compatible with all operating systems. In addition, neither Galaxy, BEDtools or SAMtools allow for the analysis of the variation of the distribution pattern of ChIP-Seq reads across a group of genes, i.e., the ChIP-Seq read density profile. ChIP-Seq read density profiles display the continuous signal across genomic regions, (e.g. around the TSS of genes) and have been widely used in the community as a way to evaluate sites of enrichment across genomic regions, and to compare profiles between different groups of genes. For instance, recent genome-wide studies on RNAPII modifications (Kim et al., 2010a, Tietjen et al., 2010, Rahl et al., 2010) have extensively explored the pattern of enrichment of several CTD modifications at gene promoters, gene bodies and genes 3' ends using this type of aggregate plots. This type of aggregate plots is extremely useful specifically in the of study broad type of data. Two applications can be used for this purpose, CEAS (Ji et al., 2006) and the repitools R package (Statham et al., 2010), but none of them offers friendly user interface like Galaxy does and are limited to pre-defined gene-related positions in the genome (when, in fact this type of profiles could be used with any genomic feature, e.g. CpG islands, enhancers, introns, exons, etc). In addition, neither Galaxy, CEAS or repitools, offer the possibility to use informative statistical tests to compare signals between groups of genomic intervals or between experiments.

In this Chapter, I describe a software, SeqGI, developed to analyse ChIP-Seq data. This software integrates visualisation with statistical testing and normalisation tools, and operates on cont-type data. Characteristics of SeqGI in comparison with other packages discussed here are presented in Table 6.1

Table 6.1: Comparison between SeqGI with available packages to visualise and analyse sequence count data.

	GUI	Read counts at genomic features	Visualisation of read density profiles	Statistical testing
Galaxy	Yes	Yes	No	No
BEDtools	No	Yes	No	No
SAMtools	No	Yes	No	No
CEAS	No	Yes	Yes (default regions)	No
Repitools (R)	No	Yes	Yes (default regions)	No
SeqGI	Yes	Yes	Yes (user defined)	Yes

6.3 SEQGI OVERVIEW AND IMPLEMENTATION

SeqGI, "**Seq**uence read enrichment at **G**enomic **I**ntervals", is an application that provides a collection of easy-to-use methods for any scientist to analyse and visualise genome-wide ChIP sequencing output data (Figure 6.1). SeqGI can be used for common research question faced at the initial steps of the analysis, for example, what are the number of reads overlapping a list of promoters? Are reads enriched at a particular group of promoters? What is the distribution pattern of a given epigenetic mark around a given genomic feature of interest (such as genes TSSs)? What is the statistical significance of differential ChIP-Seq signals between groups of genes or between samples?

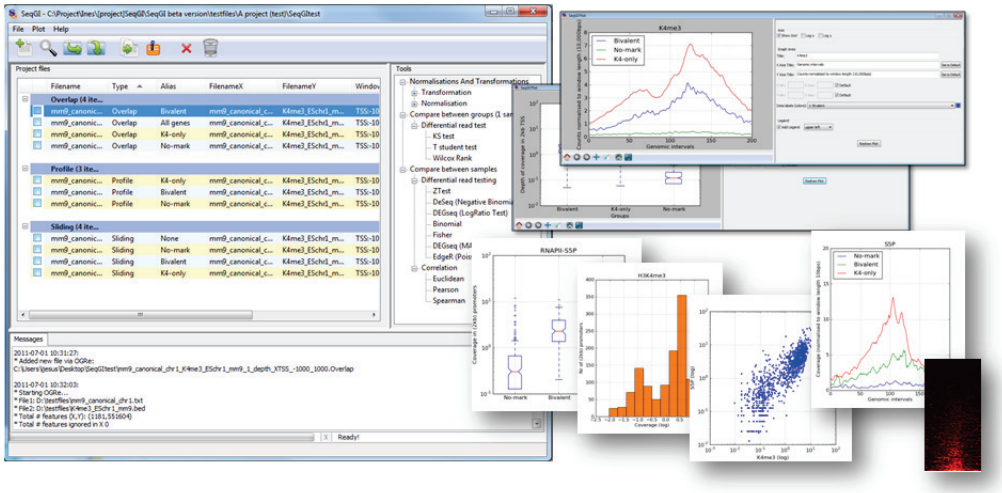


Figure 6.1: A collection of SeqGI screen shots

6.3.1 *SeqGI workflow*

The workflow of SeqGI is illustrated in Figure 6.2. It has three main components: (i) Identification of overlaps between two sets of genomic intervals, for example, a list of aligned DNA reads (feature set Y) are used to quantify the number of sequenced reads overlapping a set of genes/promoters/etc. (feature set X); (ii) graphical visualisation of the results, for instance Box plots or histograms can be used to visualise the distribution of read density between different classes of genomic features as well as between conditions, thus allowing for the identification of

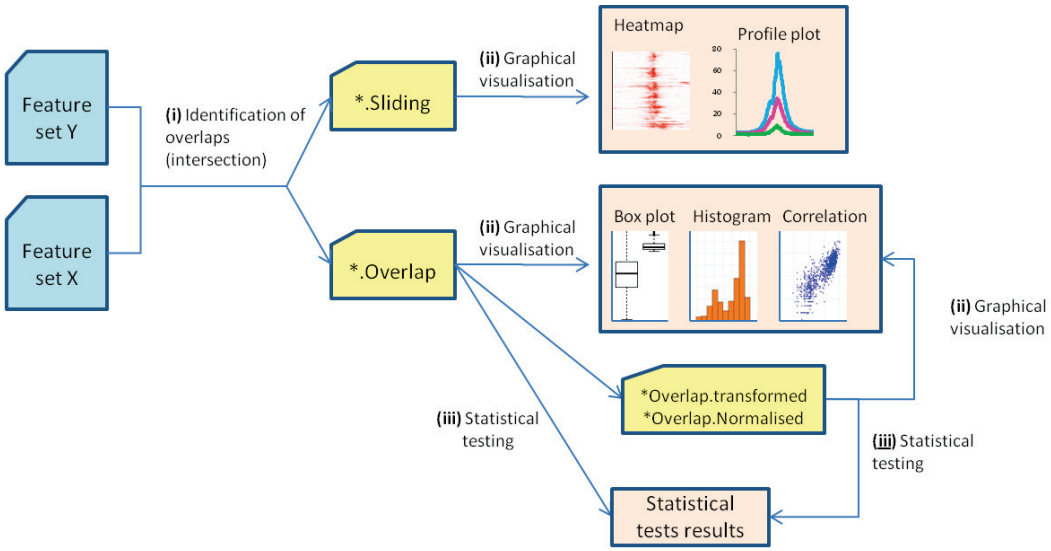


Figure 6.2: The workflow of SeqGI. Two sets of genomic intervals (X,Y) are required. This could consist for instance in a gene annotation table and a BedGraph file with ChIP enrichment signal. SeqGI has three main key features: (i) Identification of overlaps between the two sets; (ii) graphical visualisation of the results; (iii) and a statistical tools which include data transformation, normalisation and testing. Input files are displayed in blue; output files are depicted in yellow and output results in orange.

differential read densities; (iii) and a statistical toolbox to transform and normalise data and to compare count-type data between groups (single sample) and/or between samples.

6.3.2 *Layout of the GUI*

SeqGI working area is divided into three main areas (Figure 6.3):

1. On the left panel is the list control of project files, where all user project files are listed and displayed.

2. On the right panel is the statistical analysis toolbox. The user can choose from a variety of options, from normalisation, transformation, classical parametric and nonparametric tests to the analysis of differential read densities as count-type data.

3. On the bottom panel is the message log, where all the relevant messages are displayed, including workflow messages when a new calculation is running

and any error messages. All the messages are saved automatically so when the user closes SeqGI and returns to the same project later on, the relevant messages can be used to track the tasks performed previously.

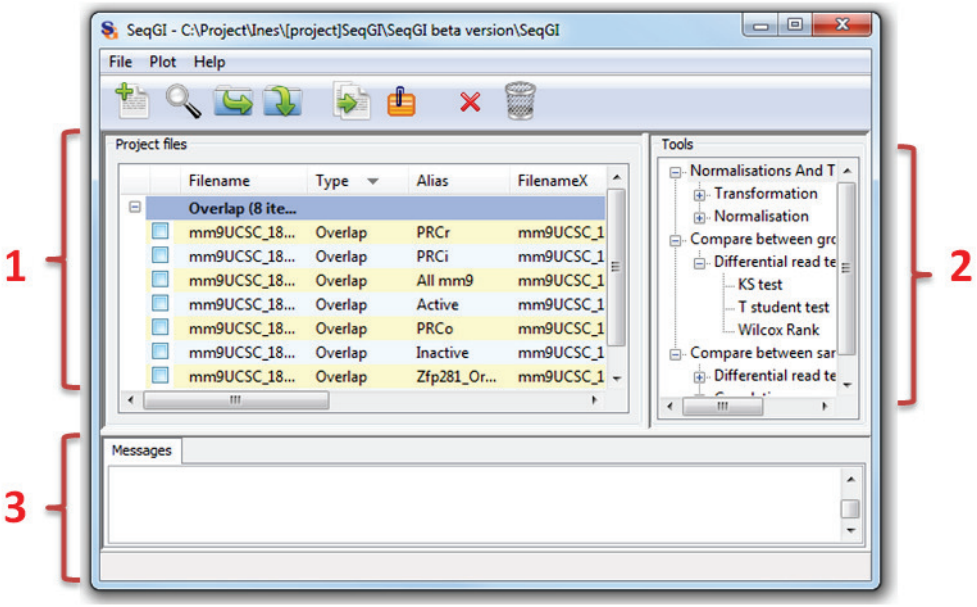


Figure 6.3: Layout of SeqGI GUI. The working area is divided into three main areas: 1. The list of project files; 2. Statistical analysis toolbox; 3. Messages panel.

6.3.3 *Identification of overlaps*

The initial step in a workflow within SeqGI is the computation of overlaps between two feature sets (X,Y). The computation of overlaps is mandatory to continue analysis under SeqGI. Typically, SeqGI would be used to calculate how many reads from a given ChIP sample overlap with each gene in the genome. Overlaps between features are calculated using OGRe script, which can also be used by more advanced users as a stand-alone script.

6.3.4 *Types of Input files*

SeqGI can be used to intersect BED, BedGraph or output files from standard aligners (feature Y) with a set of dictated genomic features (feature X). Genomic features may include for e.g. CpG islands, enhancers, peak positions, gene lists, or any list of

genomic intervals, as long as it contains *chromosome*, *start*, *end* information. Within a given gene list, gene regions of interest can be defined internally, for e.g. regions upstream/downstream TSSs or TTSs, centered at the TSS/TTS, or any region relative to gene positions. Sequencing reads are accepted in BED, BedGraph and SAM formats, or any other custom format, as long as it contains *chromosome*, *start*, *end* information for each sequenced read.

The input file types can be either "Gene list", "Peaks", "BedGraph", "BED", or a "Custom format".

Gene list and Peaks format

"Gene list" and "Peaks" are a type of file format that is specific to SeqGI. This file format is a standard text-based format with a set of coordinates that are used internally, upon user instructions, to define regions of interest. The "Gene list" file is used to define regions relative to gene positions (relative to TSS, TTS, etc), while the "Peaks" file is used to define regions relative to a peak maximum position. The "Peaks" file can be useful, for instance to visualise the distribution of sequencing reads in the vicinity of predefined peaks.

BedGraph

BedGraph is tabular format developed for use with the UCSC genome browser (Kent et al., 2002). Chromosome, start and end coordinates and a coverage value is specified in the following format:

chromosome *start* *end* *coveragevalue1*

chromosome *start* *end* *coveragevalue2*

For ChIP-Seq and RNA-Seq datasets, the coverage value corresponds to the depth of coverage in the genomic position.

BED

Bed is tabular format also developed for use with the UCSC genome browser (Kent et al., 2002). The first three fields in each feature line are required and consist of *chromosome, start,end.* Other fields are optional for the use of SeqGI.

BED3 format: *chromosome* *start* *end*

BED6 format: *chromosome* *start* *end* *ID* *score* *strand*

6.3.5 *Flexibility of input file structure*

At this step of the workflow, where the user selects two files to intersect, several options need to be considered. These include file descriptors such as column separation and column positions; file format; and whether the file is sorted or not. The possibility to specify the structure of the input file is an important characteristic of SeqGI as it confers flexibility on the type of files that can be used, for instance columns can be in any order and separated by any common column delimiters. The column delimiters can be tab, space, comma or semicolon. A preview of the first 100 lines helps the user to identify the column positions and delimiters used in the inputted file.

6.3.6 *File sorting*

It is mandatory that the genomic coordinates within the input files are sorted by start position. If the files are not sorted an option "Sort file 1/2" needs to be selected to sort the files internally before scanning the files for overlap. The internal sorting is memory intensive as it depends on the loading of the complete file into the memory. Improvements on this part of the code are currently underway. One way to go around this issue is to use previously sorted files (using command line tools such as the "sort" command in Unix). To prevent mistakes in the calculation of the overlaps, SeqGI checks if the genomic intervals are in ascending order as it is running.

6.3.7 *Intersection function*

The user can quantify the overlap between two genomic intervals (X,Y) using the sequencing depth score (i.e., each nucleotide from the read that is overlapped contributes to one count) or considering uniquely the start position of each read. Within SeqGI these two options are known as "depth of coverage" and "number of overlaps" (Figure 6.4). If sequencing reads are imputed in the BedGraph format, the intersection function available is the "depth of coverage" since the BedGraph type of files do not have single-read information, but store count-type data. The user can also specify the minimum overlap length. This setting corresponds to the minimum overlap between the two features. The default value is 1 nucleotide which is the minimum possible value.

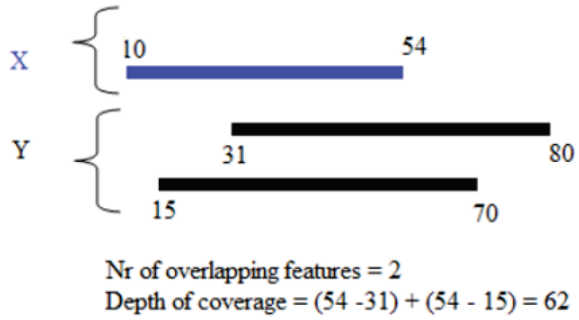


Figure 6.4: Example of intersection between two genomic intervals. The overlap can be scored using either the sequencing "depth of coverage" score or considering uniquely the start position of each read ("nr of overlapping features").

6.4 SEQGI OUTPUT AND RESULTS

6.4.1 *Overlap scores are given in two types of files "Overlap" or "Sliding"*

The user can investigate and illustrate read density at genomic intervals (or at some defined positions relative to known regions/features), by choosing between two types of outputs, "Overlap" and "Sliding". Importantly, each output file is linked to ".fields" file which contains information of the original input files and all options chosen by the user to create a given output. The "fields" file is saved internally in SeqGI project folder, and is used by SeqGI to display the file information when requested by the user. Other softwares that are commonly used in bioinformatics,

such as Galaxy (Giardine et al., 2005) or GSEA (Subramanian et al., 2005), do not allow the user to check the steps that were taken to create a given project file. From the user's perspective, it is important to maintain record of all steps taken to create a given output, to allow for checks and to prevent future mistakes or the need to repeat the whole analysis.

Overlap

In the "Overlap" type of output, the overlap between feature X and Y is calculated by using the total length of feature X; for example this option can be chosen to calculate the depth of coverage at the ±1kb region around the TSS.

The format of "Overlap" files contain 7 tab delimited columns:

- ID, chromosome, start, end, strand derived from file 1 (feature set X)
- Window length: size of the window (end – start)
- Counts: calculated as either the number of overlaps or depth of coverage
- Average counts: counts normalised to the window length

Sliding

The "Sliding" format, allows the user to explore the pattern of the ChIP signal at the genomic intervals of feature-set X. The distribution profile is calculated by using a sliding window that scans each genomic interval defined in feature-set X. If the genomic intervals in feature-set X contain *strand* information (applicable when feature X is a gene or contains gene-related coordinates), then the sliding windows are defined for all genomic intervals in the same orientation (5′ to 3′), otherwise left to right would be used. The length of the sliding window can be either a fixed size (in bps) or a percentage of the total size of feature X. Defining the sliding window as a percentage is particularly important when the genomic intervals are not all of the same size, thus the percentage is used to scale each interval to a common length.

The format of "Sliding" files consist of:

- Five initial feature descriptors (as in *.Overlap): ID, chromosome, start, end, strand derived from file 1 (feature set X)
- Columns after the 7th column correspond to counts (either the number of overlaps or depth of coverage) normalised to the window length.

6.4.2 *Graphical visualisation*

Once the overlap scores have been determined, the distribution of scores can be visualised both between different sets of genomic intervals, or between subgroups of a given set of genomic intervals. Graphical representations available include profile plots, heatmaps, Box plots, histograms and scatter plots.

- **Profile plots** and **heatmaps** are used to visualise the pattern of read distribution across features. Both plots are created from the "Sliding" type of files. The profile plot averages the signal across all the members of a selected group of features (e.g. group of genes), while the heatmap displays the two-dimensional data as colors, and thus does not depend on the average signal per group.
- **Correlation plots** (scatter plots) are used to compare two "Overlap" files, for instance to visualise the degree of correlation between two ChIP samples. This plot can be complemented with the use of a statistical analysis such as Spearman's rank or Pearson correlation tests.
- **Box plots** and **histograms** can be used to graphically depict the distribution of data within "Overlap" files. Box plots use the standard descriptive statistics: the minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and the maximum. Box plots are a convenient way to compare the ChIP signal across groups. These can be complemented with the use of hypothesis tests within SeqGI, such as Wilcoxon rank-sum test or Student's t-test.

6.4.3 *Statistical testing*

Coupled to SeqGI's rich visualisation tools, SeqGI implements a core set of normalisation, transformation and sequence read enrichment testing procedures. SeqGI's statistical framework can be used to answer questions such as, is a given group of genes more enriched for a particular protein/modification then another group? What are the genes that show differential occupancy between two given conditions (e.g. wild-type *versus* knockout, undifferentiated *versus* differentiated)? Are two datasets correlated across gene promoters? Alongside the analysis of differential read densities, data transformation and data normalisation can also be applied to the datasets. For example, data normalisation can be used before comparing data scoring between two samples; SeqGI offers several options for data normalisation including quantile normalisation, RPKM, median scaling, and others. The code for the implementation of the statistical tools in SeqGI is written in R and is currently being developed in collaboration with Tom Carroll.

- **Transformation:** Genomic Intervals can be rapidly normalised to interval lengths as well as library sizes and transformed from raw count data to Log_2, Log_{10} or arcsine values for conventional parametric testing. The transformed files are then available from the SeqGI project to allow for further visualisation and testing, according to the user's requirements.

- **Normalisation**: Normalisation options include, Reads per million (RPM), Reads per kilo-base per million (RPKM), Quantile normalisation, Variance Stabilizing Normalisation and Median Scaling.

- **Differential read tests - Between Groups:** SeqGI implements a selection of testing procedures established in SAGE, RNA and ChIP sequencing analysis. Groups can be quickly compared in SeqGI using both parametric (Student's t-test) and non-parametric tests (Wilcoxon rank-sum test).

- **Differential read tests - Between Samples:** SeqGI's statistical framework allows for the analysis of differential read densities as count data using

methodology implemented in the DESeq (Anders, 2010) and DEGseq (Wang et al., 2010) Bioconductor packages.

- **Correlation between samples:** Correlation over genomic intervals between conditions may be evaluated using Pearson or Spearman's correlation tests.

6.4.4 *Implementation*

SeqGI is written in Python and R. SeqGI runs in Windows and is being currently developed to run in other standard operating systems (MAC, Linux).

6.5 COMPUTING OVERLAPS BETWEEN GENOMIC INTERVALS - OGRE SCRIPT

OGRe ("**O**verlap **G**enomic **Re**gions") is a stand-alone script that was developed to be used within SeqGI to find overlaps between two sets of genomic intervals. OGRe is a major feature of SeqGI as it allows for fast processing of the analysis. It can also be used independently of SeqGI as command line tool. The basis of OGRe algorithm is an adaptation of the original algorithm "Fjoin" (Richardson, 2006).

6.5.1 *Algorithms to compute the overlap between two genomic feature sets*

Given two feature sets (X,Y) where each interval is defined by a set of coordinates (*chromosome*, *start*, *end*), Fjoin (and OGRe) finds which intervals in the Y set overlap the intervals in the X set.

A straight forward algorithm that can be used to answer this question is the nested loop. In this algorithm, two loops are used, one loop over all features of set X and a second loop over all features in set Y (Richardson, 2006). However, for large genomic datasets, the use of nested loops becomes impractical since a loop between all Y features is required for each feature in set X. Some practical approaches can be applied to reduce the time needed to iterate through all possibilities, for instance if the lists X and Y are previously sorted by start positions, one can break the loop in file Y as soon as the start position of the genomic interval Y is bigger then the end position of the genomic interval X. Although this approach reduces the time needed

to iterate through all pairwise possibilities, it is still very time consuming to perform this type of search in large genomic datasets. One way to overcome this issue is by the use of interval trees (a data structure originally proposed by Edelsbrunner and Maurer, 1981), which mainly consist of an efficient data structure that, once built, can be quickly used to make fast queries. Although interval trees are a good option in a database-like environment, in many occasions it is preferable to use a simple command line tool that does not depend on the pre-installation of a database system and that can be used without having to go through the CPU time expensive tree construction stage.

6.5.2 *Fjoin "sort merge-pass" algorithm*

The solution to this problem, used in Fjoin, is the use of a "merge-pass" algorithm. This algorithm makes use of sorted lists of the starts and ends of all the intervals and scans every interval from left to right, more or less in parallel, outputting when an overlap is found. The "merge-pass" is basically a sliding window algorithm. The algorithm operates on two iterators of intervals (that scan X and Y sets). In each round, the genomic intervals in a given set are scanned against a genomic interval in the opposite set.

For instance, consider that we are scanning a set of Y features against a single interval ("bait") in set X. Features in set Y are scanned one by one, and if located at the left of X they are dropped, while others are "caught" until a Y feature is found at the right of X. Then the order of the scan is inverted, *i.e.*, in parallel with the previous round, features of set X are scanned against a "bait" feature Y, until the search passes the left interval. Then the order is reversed again and the previously "caught" Y features are scanned against a feature X, and so on. The Fjoin algorithm allows for the rapid search of overlapping features in a remarkably simple procedure (Richardson, 2006).

6.5.3 *OGRe improvements on Fjoin*

I used the original algorithm in Fjoin to build OGRe. Several key points needed to be adapted from the original algorithm to allow its use with large high-throughput

genomic datasets and for with specific data formats required for downstream analysis and visualisations.

Improvements on the original algorithm are the following:

1. In OGRe, the user can choose the intersection function as either "depth of coverage" or "number of overlaps", as explained above in subsection 6.3.7. In Fjoin algorithm, every overlapping pair is outputted as soon as the overlap is found, and thus every feature in the set X is repeated as many times as an overlap with set Y is found. This type of output revealed to be impractical for high throughput type of data, where hundreds or thousands of reads may overlap a given interval (for instance gene bodies can be several kb long and thus many reads of 32-150 nucleotides may overlap). In OGRe, the overlap is outputted only in relation to the features in set X, in a way that each feature X is written only once in the output file, with the correspondent overall overlap score found.

2. The overlap score is recorded for all features in set X, independently of an overlap being found or not. This was not possible in Fjoin algorithm. In Fjoin, features in set X for which no overlap is found are not recorded in the output. In the case of ChIP-Seq data, it is essential to report the genomic features for which no overlapped read is found. I implemented this step of the algorithm by outputting the genomic interval X and its correspondent overlapping score at the step where the item would be passed according to the original Fjoin algorithm. This is the critical step to record the overlapping score because that particular feature X is not going to be used in further scans.

3. OGRe supports BED and BedGraph files (SAM and MAP file support is currently being developed). This implementation was added to the original algorithm by simply including a description of the correspondent default column positions.

4. OGRe allows the user to choose a window of interest from a "Gene list" or "Peaks" file. When the user specifically defines the input file format as a "Gene

list" or "Peaks", the option to define a window of interest is enabled. The window of interest is characterized by coordinates which are defined in relation to genes (e.g.: 2kb upstream TSSs, Gene bodies, etc) or in relation to peak maximum position (e.g. ±1kb surrounding the peak maximum).

5. OGRe is also able to calculate the overlap pattern in each genomic interval of set X. This is accomplished by binning the genomic intervals of set X in smaller windows of a given fixed length or of a given percentage. After a overlap between a X,Y pair is found, each sub-interval in X is scored according to the intersection function chosen.

6.6 SAMPLE WORKFLOW OF SEQGI

6.6.1 *Installation and use*

SeqGI source code can be downloaded from the project web page (HTTP://seqgi.sf.net). Dependencies include Python itself, wxPython, NumPy, matplotlib and ObjectListView. R is currently distributed with SeqGI and thus does not require previous installation. NumPy and matplotlib are only required for visualisation operations, and therefore can be skipped if the user does not wish to use these tools.

6.6.2 *Adding new files to a project*

As an example, we can consider comparison of the ChIP S5P signal obtained in promoters of bivalent genes vs 20% most and least expressed genes (top20, bottom20, respectively, as in Chapter 3). A typical workflow would be (Figure 6.5):

1. Intersect a list of all genes (file 1) with the S5P aligned reads (e.g. BedGraph format; file 2)
2. Specify a region of interest in file 1, for instance ±1kb from TSS
3. Choose "depth of coverage" as the intersection function

4. Choose “Overlap” or “Sliding” as the type of output file

5. Group files using the grouping option. Within SeqGI, genomic features can be quickly characterised into pre-defined groups using the GMX standard format. In this example the GMX file contains “bivalent”, “top20” and “bottom20” groups.

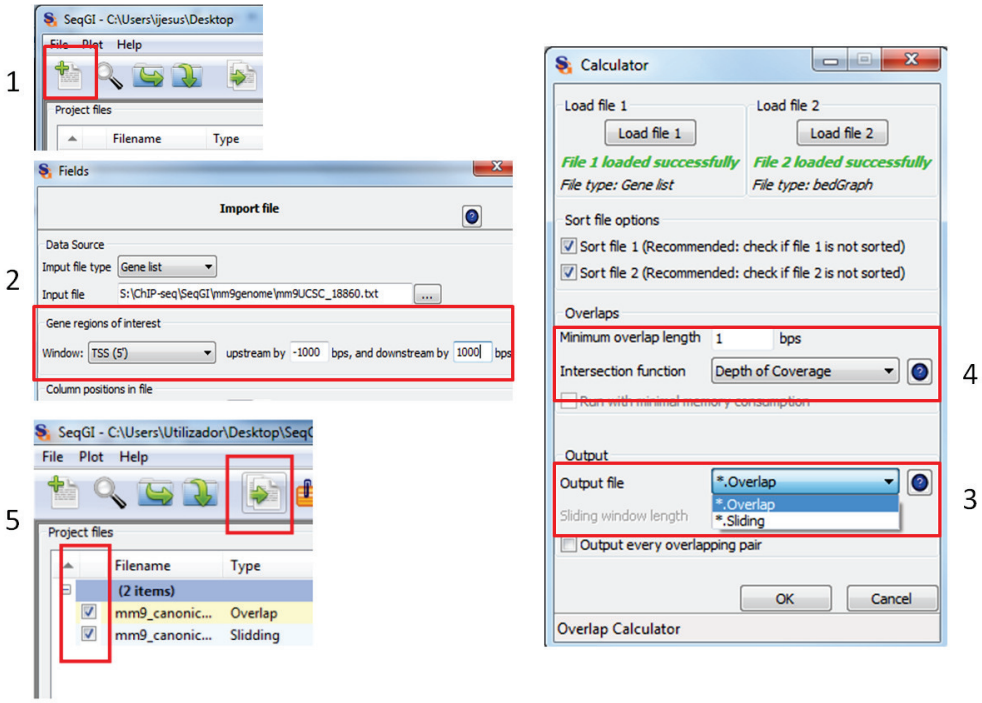


Figure 6.5: Sample workflow of SeqGI - Adding new files to a project

6.6.3 *Graphical representation*

As an example, we can visualise the differential read distribution between the different groups of genes using “Box Plots” on the “Overlap” files and “Profile plots” on the “Sliding” files. This can be accomplished by simply choosing the “Overlap”/”Sliding” file and then the correct plot from the Menu (Figure 6.6). The visual analysis using Box plots can be complemented with the use of statistical hypothesis tests such as Wilcoxon rank-sum test or Student’s t-test.

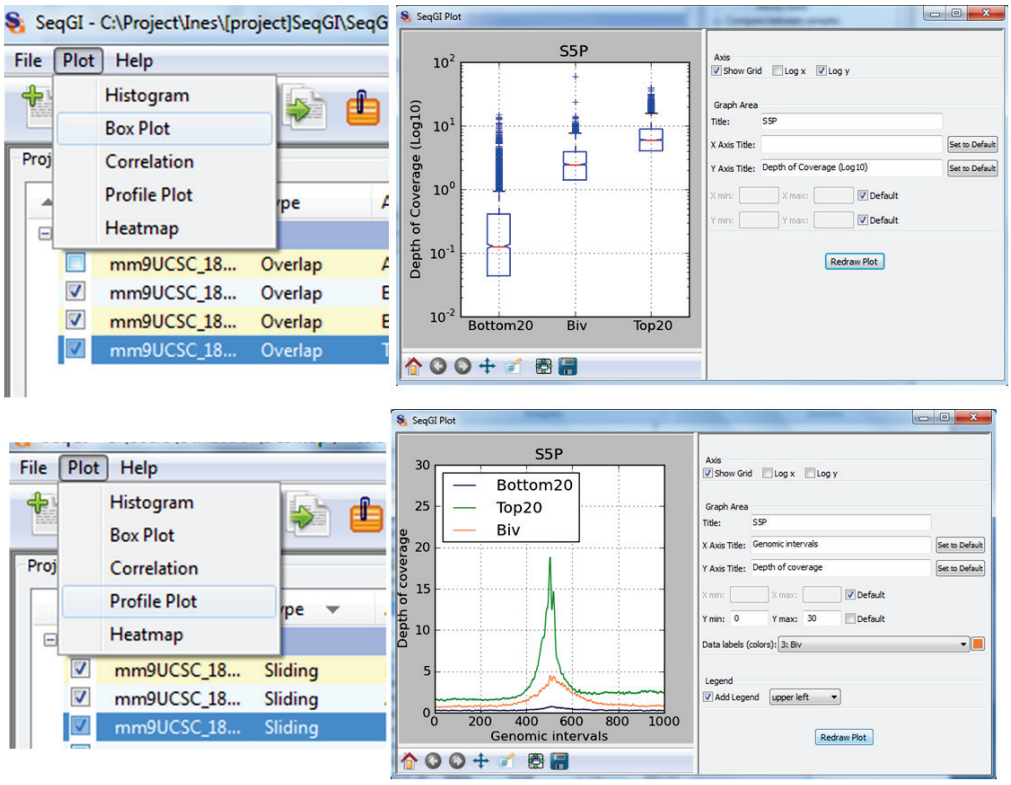


Figure 6.6: Sample workflow of SeqGI - Example plotting a Box Plot

Alternatively, it is also possible to plot a transformed dataset. To do so, a transformation function can be chosen from the statistical tools (e.g. “log 2”) to transform an “Overlap” file and then display the transformed results again using Box plots.

6.7 CONCLUSION

Understanding the mechanisms of gene regulation is a major aim of genome-wide mapping studies using next generation sequencing. ChIP-Seq technologies provide a detailed characterisation of various epigenetic features and binding patterns across different genomic regions at unprecedented resolution and dynamic range but the examination of genome-wide data requires the use of several bioinformatic tools, from the alignment of reads to a reference genome to extensive downstream analyses of binding patterns on a genome-wide scale. The analysis of epigenetic states has revealed common epigenetic profiles that are related to regulatory functions and repeated throughout the genome. For example, active genes are associated with H3K4me3 at gene promoters and H3K36me3 through coding regions, whereas

bivalency of H3K4me3 and H3K27me3 at gene promoters reflects Polycomb regulation and gene poising for expression upon activation (Mikkelsen et al., 2007). Another striking example is the finding that H3K36me3 is enriched at the exons of active genes, revealing an interplay between chromatin modifications and the splicing machinery (Kolasinska-Zwierz et al., 2009). Therefore, it is important to consider and quantify these different combinations and distinct profiles of ChIP signal that exist on genes as it can have important implications for gene regulation. In the context of high-throughput genome analysis, few applications allow for the investigation of ChIP signal as read densities at defined genomic intervals of interest. Packages such as BEDtools (Quinlan and Hall, 2010) and SAMtools (Li et al., 2009) enable a number of operations using BED and SAM type of files, including the computation of overlaps between a set of reads and a set of genomic features, however both depend on the use of command line and more advance bioinformatic skills. Other tools such as CEAS (Ji et al., 2006) and the repitools R package (Statham et al., 2010) can be used for the computation of the profiles of read density on genes, but provide the user with limited user-friendly interface and few options for statistical testing. In any case, a comprehensive output for graphical interpretation and visualisation of the results from this type of analysis of count-type data was not yet combined in one single application.

SeqGI is an open source software that provides users with an intuitive graphical interface for the simultaneous visualisation, grouping and testing of sequence read distributions both between and within classes of user defined genomic features. It can create plots to visualise the enrichments of a given mark along gene bodies or the average enrichment for a given set of regions in one or different conditions. Along side this, it provides a wide range of statistical tools that can be used. SeqGI can be used in a variety of downstream analyses integrating hypothesis drawn from visual and statistical observations and providing a platform for further biological questions to be asked.

Part IV

DISCUSSION

7

DISCUSSION

7.1 THESIS OVERVIEW

Polycomb repressor complexes (PRCs) are important regulators of gene expression implicated in pluripotency and cancer (Richly et al., 2010). A previous study of nine PRC targets in my group had found that Polycomb silencing in ES cells co-exists with active chromatin and an unusual form of RNAPII (Stock et al., 2007), but the relationship between PRCs and RNAPII genome-wide remained unclear. In this thesis I have investigated the interplay between RNAPII and Polycomb repressor complexes (PRCs) at genes in mouse ES cells, using ChIP-Seq and mRNA-Seq.

Clear chromatin and RNAPII patterns were identified at three major states in the ES cell genome: Active, Inactive and PRC-repressed (Figure 7.1 and Table 7.1).

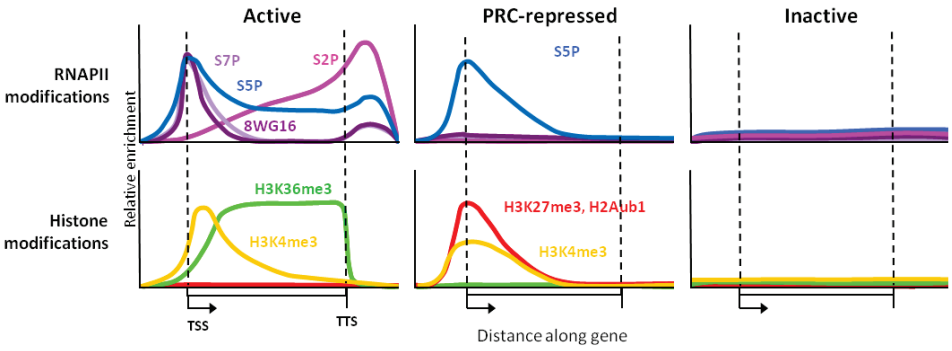


Figure 7.1: Schematic representation of average RNAPII and histone modification profiles at "Active", "PRC-repressed" and "Inactive" states. Main characteristics of each group are summarised in Table 7.1. This diagram was adapted from Brookes [2010].

Through classification of RNAPII and histone modification patterns, followed by clustering, four subpopulations of PRC-repressed genes were found (Figure 5.7): the "true" PRC-repressed state (namely, PRCr class) consisting of PRC-repressed genes occupied by unproductive (S5P) RNAPII; and three intermediate classes consisting of genes with shared characteristics of both "Active" and "PRCr" (namely PRCi and PRCa genes), or both "PRCr" and "Inactive" (namely, PRCo genes). Genome-wide analyses, together with sequential ChIP at single genes, show that PRCs co-associate with RNAPII-S5P complexes at promoters and throughout coding regions but not with RNAPII-S2P. The existence of "active" PRC-targets suggested that a subpopulation of PRC-repressed genes switches between 'on-off' chromatin states that exist either at different alleles of the same gene or across the ES cell population, potentially in response to fluctuations in levels of pluripotency transcription factors.

Although much remains to be explored about the way specific CTD and epigenetic states are "read", our results show that specific combinatorial patterns are established at differentially regulated genes, suggesting that regulation of CTD phosphorylation is important for robust regulation of gene expression and genome function.

Table 7.1: Summary of the three major epigenetic states in ES cells: "Active", "PRC-repressed" and "Inactive". RNAPII modification correlates with expression outcomes and chromatin state.

	Active	PRC-repressed	Inactive
mRNA expression	High	Low	Undetected
RNAPII recruitment	Yes	Yes	No
RNAPII (S5P) and H3K4me3	High S5P (promoters and coding regions). High H3K4me3 (promoters)	High S5P (promoters and coding regions). High H3K4me3 (promoters)	Undetected
RNAPII (8WG16)	High (promoters)	Low or absent	Undetected
RNAPII (S7P)	High (promoters and coding regions)	Low or absent	Undetected
RNAPII (S2P) and H3K36me3	High S2P (coding regions and downstream 3′ end). High H3K36me3 (coding regions)	Low or absent	Undetected
PRCs	Low or absent	High (promoters and coding regions). Proportional enrichment to S5P.	Low or absent
Promoters	CpG-rich	CpG-rich	CpG-poor
DNA methylated promoters	Not enriched	Not enriched	Enriched
Gene Ontology	Mainly housekeeping functions such as "gene expression", "DNA repair" and "metabolic function"	Developmental regulator and metabolic functions	Specialised adult-related functions such as "multicellular organismal process"

7.2 FROM RESEARCH OBJECTIVES TO RESEARCH FINDINGS

7.2.1 *Bioinformatic analysis and software development*

High throughput ChIP-Seq and mRNA-seq methods have provide the research community with the power of quickly acquiring genome-scale maps of a variety of

factors. The integration of results from different levels (epigenomics, transcriptomics, proteomics) and experiments are challenging but offer valuable and unprecedented insight into the genome function and regulation.

To investigate the logic of complex genome regulation in ES cells, I have integrated and mined the results from a variety of studies. I used ChIP-Seq and mRNA-Seq methodologies to generate genome-wide maps of RNAPII CTD phosphorylation patterns at S5, S7 and S2 residues, PRC1 subunit (Ring1B) and two histone modification marks associated with PRC-mediated gene repression (H2Aub1) and with transcription elongation (H3K36me3). In addition, along side our newly generated maps, I have integrated several published datasets to investigate RNAPII modification status and its relationship with characteristic epigenetic signatures across the ES cell genome. Published datasets included ChIP-Seq (Ezh2, Suz12, core H3, H3K4me3, H3K27me3; Mikkelsen et al., 2007, Ku et al., 2008); mRNA-Seq (Cloonan et al., 2008); Microarray gene expression in ES cells (Mikkelsen et al., 2007, Sene et al., 2007), during ES cell differentiation (Shen et al., 2009) and after Ring1A/B dKO (Endoh et al., 2008), and other ES cell genome-wide annotation datasets, namely promoter DNA methylation (Fouse et al., 2008), proteome characterisation (Graumann et al., 2008, Van Hoof et al., 2006) and TF binding (Kim et al., 2008). Integration of data sets from a variety of sources is an essential part of data analysis. To this end, I have used several bioinformatic softwares that were already publicly available, such as GSEA (Subramanian et al., 2005) and MultiExperiment Viewer (Saeed et al., 2003). Additionally, I used R and Python to generate my own scripts and analysis workflows.

New NGS technologies generate large amounts of data to process and analyse. This has become a challenge to the bench scientist (not to mention the bioinformatician itself). In the last year of my PhD, I have used wxPython, Python and R to develop a software, SeqGI (Chapter 6), that can assist on the first steps of analysis and interpretation of high-throughput data. SeqGI can be used to visualise known gene profiles of a given set of factors (such as the ones investigated in this thesis, CTD and histone modifications) and to compare the signatures at pre-defined gene groups or genic regions (for instance, highly expressed versus lowly expressed genes, exons

versus introns, short versus long genes, etc). It provides the user with an intuitive graphical interface combining both visualisation and statistical tools, and thus allowing any scientist to quickly take initial steps in complex data analysis. Along with SeqGI, I have developed OGRe, which can be used to quickly calculate read density at any genomic interval or genic regions of interest. Result files produced by OGRe are integrated within the analysis performed by SeqGI (visualisation, statistical testing) but can also be used independently to suit appropriate analysis needed to answer particular questions.

Although analysis of sequencing data using SeqGI can be highly informative, more advanced data mining techniques are often needed. For instance, in this thesis I have used hierarchical clustering (Chapter 4) and principal component analysis (Chapter 5) to uncover relationships and correlations between RNAPII and Polycomb-instigated histone modifications within the data. Ultimately, it is important to mention that, despite the power of bioinformatics, data analysis should be coupled with experimental validation as often as possible.

7.2.2 *A review of the CTD central dogma*

The current view of the RNAPII CTD code applies generally to the results described in this thesis, namely that S5P and S7P are enriched at the 5′ end, while S2P levels increase during elongation. However, a number of interesting examples of unknown combinations of CTD phosphopatterns was revealed in the course of this study.

For instance, highly transcribed genes exhibit S5P and S7P enrichments at the 3′ end (Figure 3.7), despite the current idea that S5P is removed by protein phosphatases during transcription elongation (Mosley et al., 2009). Accumulation of RNAPII at the 3′ end may reflect RNAPII slower kinetics probably important for correct mRNA processing. Furthermore, the peak of S2P downstream of the 3′ end is not accompanied by increased levels of H3K36me3, suggesting that the interaction between HMTs and S2P is disrupted at this stage.

Other interesting examples come from the characterisation of an unusual RNAPII configuration at PRC-repressed genes. For instance, we have identified that many of

these genes harbor S5P throughout their coding regions but S7P and S2P are low or undetected (Figure 3.9 and Figure 5.19). Considering that the kinase Cdk7 is responsible for the phosphorylation of both S5 and S7 residues of the CTD (Akhtar et al., 2009, Glover-Cutter et al., 2009), S5P in the absence of S7P at PRC-repressed genes suggests that specific S5P kinases or specific S7P phosphatases may exist. S2P is important for correct mRNA processing and recruitment of splicing and polyadenylation machinery (Hirose and Ohkuma, 2007). Similarly to our own findings in mouse, S7P was found at coding regions in humans (Chapman et al., 2007) and yeast (Tietjen et al., 2010, Kim et al., 2010a, Mayer et al., 2010), and it was suggested to have a role in elongation and co-transcriptional splicing. Detection of appreciable amounts of S5P (and H3K4me3 at the promoters) in the absence of S2P and S7P (and H3K36me3) at PRC-repressed genes suggests that, S5P in the absence of S2P is congruent with elongation. In this context, while RNAs are produced (Stock et al., 2007), efficient RNA processing is likely to be disrupted.

7.2.3 *Interplay between RNAPII and Polycomb at PRC-repressed genes*

In this study I have identified four classes of PRC-repressed genes: Three groups displayed CpG-rich promoters, PRC and RNAPII binding (PRCr, PRCi and PRCa). Levels of mRNA and marks of actively transcribing RNAPII (S7P, S2P) were absent at PRCr genes but levels increased from PRCi (S5P+S7P+) to PRCa (S5P+S7P+S2P+) genes. We suggest that PRCr class constitutes the "true" PRC-repressed state, while PRCi and PRCa genes fall on a sliding scale of PRC-regulated to active states. Additionally, we identify a cohort of PRC-repressed genes (PRCo class) which do not recruit RNAPII, have CpG-poor promoters and seem to be regulated mostly by PRC2 (demonstrated by detectable levels of H3K27me3), but not PRC1 (as H2Aub1 levels are low or absent at the majority of genes within this group; Figure 5.7). At PRCo genes, the coincidental lack of RNAPII recruitment and H2Aub1 modification suggests that the interplay between RNAPII and PRCs may depend on PRC1 directly, or indirectly through the deposition of H2Aub1 marks and the recruitment of appropriate regulatory factors.

An unusual pattern of CTD modifications was found at PRCr genes (Figure 3.9 and Figure 5.19), which consisted of high levels of S5P, no detectable S2P, and an unexpected lack of detection of unphosphorylated S2 residues (recognised by antibody 8WG16). Even in the absence of S7P and S2P, RNAPII (S5P) is found at the gene bodies of many PRCr genes, and producing low levels of transcripts (Stock et al. 2007), indicating that Polycomb repression acts downstream of transcription initiation and that the absence of S2P and S7P may limit complete mRNA maturation (Stock et al., 2007, Brookes and Pombo, 2009). Yet, the pattern of S5P enrichment is different among PRCr genes. At some genes the phosphorylation mark is found more localised at the promoter region, while in others it is found through the whole coding region up to the TTS, but not beyond (Figure 5.13). In all cases, the profile of S5P is mirrored by H2Aub1 and H3K27me3 suggesting that PRCs travel through genes coincidental (in abundance and distribution) with elongating RNAPII-S5P complexes (Figure 5.12). Interdependence of RNAPII and PRCs in the PRC-repressed state was further shown by sequential ChIP, which detected enrichment for RNAPII-S5P with Ring1B or Ezh2 at all PRCr genes tested (Figure 5.14) and by bioinformatic analyses of enrichment levels at TSS, which showed strong positive correlation between S5P, H2Aub1 and H3K27me3 levels within the PRCr group (Figure 5.22).

The mechanisms controlling RNAPII extension through coding regions of PRC-bound genes are not known and merit further investigation. Underlying sequence elements at these genes may be relevant, for instance PRCr,i,a classes recruit RNAPII and contain CpG-rich sequences at their promoters, while promoters of PRCo genes are mainly CpG-poor and do not recruit RNAPII. The average profiles of S5P at PRC-repressed genes show a broader S5P peak (de-localised ~200 bp from the promoter; Figure 5.12) when compared to active genes, which could be reflecting the variable degrees of S5P extension into coding regions of these genes, or alternative TSS usage, or both. Interestingly, coincidentally with broad S5P distributions at PRC-repressed genes, we observe broad (unmethylated) CpG density around the TSS (Figure 5.11), which in turn have been associated with multiple TSS usage (Sandelin et al. 2007) and RNAPII recruitment (Rozenberg et al., 2008). Another interesting observation is that bivalency (Akkers et al., 2009) or presence of RNAPII at PRC targets (Schwartz et al., 2010) were not found in lower eukaryotes, although the

cellular systems might not be the exactly representative of the ES cell state. Interestingly, lower CpG content (FitzGerald et al., 2006) and reduced CTD complexity (Chapman et al., 2008) are also characteristics of lower eukaryotes, thus raising the possibility that the poised state of RNAPII at PRC-bound genes evolved recently, possibly in parallel to the evolution of CpG islands, and may be specific to vertebrates.

7.2.4 *Fluctuations between the canonical active and PRC-repressed states*

Cohorts of Polycomb targets (PRCi and PRCa classes) shared characteristics with PRCr genes (high levels of S5P and PRC repressive marks), but also displayed characteristics associated with the Active group of genes (high 8WG16 and S7P enrichment at promoters and S2P increasing levels through the gene body and peaking immediately downstream of the TTS). Sequential ChIP experiments showed that, at PRCi and PRCa genes, Polycomb co-localises with the RNAPII-S5P, but not with active (S2P) RNAPII (Figure 5.21), suggesting that the concomitant detection of PRCs and actively transcribing RNAPII is due to expression from these PRC-targets in a small minority of cells, or even within different alleles of the same cell. Therefore we propose that the active signature of PRC target genes is a read-out of the heterogeneous ES cell population and the existence of two subpopulations, one in which the gene is transcribing, other in which is PRC repressed. Furthermore, PRCa genes produce not only mRNA but also protein (Figure 5.8), and developmental and metabolic processes are enriched within this group, suggesting that regulation of these genes is likely to be important for ES cell biology. Several of the genes, such as *Hmga2*, have known roles in pluripotency (Pfannkuche et al., 2009).

Recent studies have shown that pluripotent ES cells can present a variety of inter-convertible states responsive to fluctuations in the pluripotency transcription factor network (Canham et al., 2010, Graf and Stadtfeld, 2008). Particularly, Nanog (Chambers et al., 2007), among other factors such as Rex1 (Toyooka et al., 2008) and Stella (Hayashi et al., 2008), is known to have fluctuating expression levels within the ES cell population thus contributing to ES cell heterogeneity. Nanog-target genes were found particularly enriched at PRCa class. However, silent PRC-repressed genes

are also known targets of Nanog (and Oct4 and Sox2) in which Nanog is thought to act as a a repressor (Boyer et al., 2005, Loh et al., 2006). It was recently suggested that TFs, such as Nanog, can function as weak activators when bound together with few other factors, but the concerted binding of multiple TFs results in gene activation (Ouyang et al., 2009, Kim et al., 2008). If we apply the same principle to the PRC-repression in ES cells, one could speculate that a "PRCa" gene may switch from a PRC-repressed state to an active state in response to the combinatorial binding of several TFs (Figure 7.2).

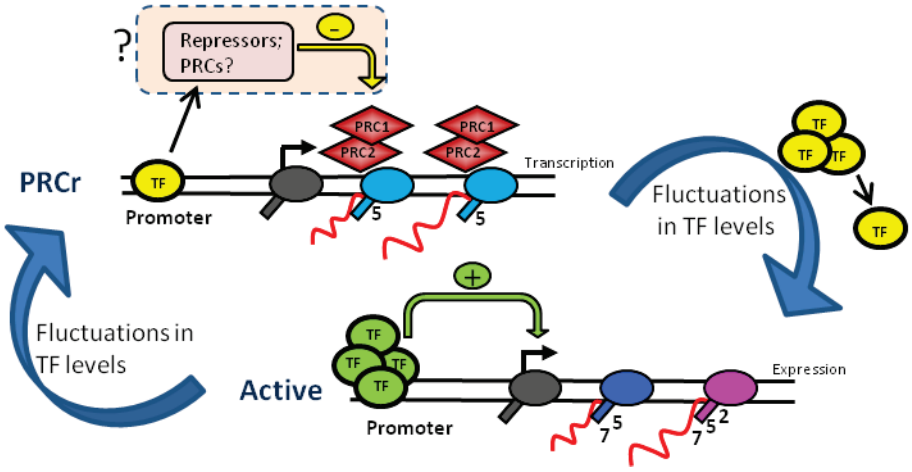


Figure 7.2: Model of the switch between PRC-repressed and the Active states in ES cells. The switch between the PRC-repressed (PRCr) and the active state is dependent on the cooperative binding of multiple TFs. Active genes are associated with S5P (blue) during initiation and S2P (pink) during elongation. PRCr genes recruit RNAPII-S5P (blue) which extends into coding regions without S2P. A TF (yellow) may bind to Polycomb repressed targets and function as a weak activator, or even act to repress its targets (as suggested by Wang et al., 2006); or it may act cooperatively with other TFs (green) to trigger gene activation.

7.3 FUTURE RESEARCH DIRECTIONS

In this study, we focused on how different combinations of CTD and histone modifications are related to PRC repression. Apart from the modifications studied in this thesis, many other different combinations of phosphorylations or other modifications can take place and could be potentially crucial to modulate gene expression states, for instance through the recruitment of chromatin modifiers or

other regulators of transcription. Additionally, other processes may be relevant in this context, such as alternative promoter usage and antisense transcription. A genome-wide scale study of nascent RNA transcripts is currently ongoing in the laboratory, both by CAGE analyses and by isolation of nascent RNA associated with chromatin-bound RNAPII. The analysis of nascent RNA transcripts together with integration with DNA sequence data and RNAPII ChIP-Seq data will be important in the characterisation of transcripts from PRC-repressed genes (if they are capped, spliced, polyadenylated, truncated or completely mature), the transcriptional machinery involved in their production, and in elucidating sequence features (such as alternative initiation sites or longer introns) implicated in this phenomena.

Our deep representation of the CTD code and transcriptome are valuable resources for the study of gene expression control in ES cells. To this point, limited data is available to analyse RNAPII and epigenetic signatures in other cell types. Mapping the genome-wide distribution of RNAPII phosphoforms in neuronal-committed cells is currently under way. Considering the major influence of Polycomb proteins in gene expression, it is not surprising that disruption of Polycomb function results in disease. Particularly, it has been shown that many Polycomb group proteins are involved in cancer (Richly et al., 2010). Ultimately, a comparison of RNAPII and Polycomb function between pluripotent, differentiated and disease states would not only contribute to our understanding of gene regulation at different stages of the cell, but it would be also of fundamental importance in order to understand and manipulate gene expression control. In addition, genomic maps of total RNAPII binding (for instance, using an antibody that targets RPB1 independently of CTD phosphorylation) would bring valuable insights about the RNAPII abundance and, by comparison with phosphorylation profiles, the levels of CTD phosphorylation at each genomic site.

The extent to which measurements in cell populations might mask the activity of different subnetworks in subpopulations of cells has recently been addressed in ES cells (Trott et al., 2012). Moreover, it was suggested that heterogeneity among protein complexes could enable a diversification of potential functions of the cellular machinery (Gavin et al., 2006). Although the idea of heterogeneous subpopulations of

interconverting ES cells fits with our observations at PRCa genes, it would also be interesting to investigate whether the composition of PRC complexes differs to enable multiple functionality. For example, a recent genome-wide study in mouse myoblasts provided evidence for the association of Ezh1 with RNAPII and active transcription (Mousavi et al., 2011). We are currently planning experiments to map Ezh1 genome-wide in ES OS25 cells in order to explore its co-association with RNAPII and Polycomb complexes at PRC-regulated genes.

Other interesting clues on the complexity of epigenetic regulatory networks and how its regulation contributes to cell phenotype diversity are quickly emerging. A particular example comes from the discovery of TET1, an enzyme found to catalyze the oxidation of 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC) in mammals (Tahiliani et al., 2009). In ES cells, 5hmC can either be associated with active promoters and gene expression, as well as with silenced PRC target genes (Wu et al., 2011). It would be interesting to explore if differential TET1 or 5hmC levels are present at different classes of PRC targets (from PRCr to PRCa genes) and thus could reflect differential methylation dynamics at these genes. In addition, we are currently performing a proteome study coupled with chromatin IP (a technique named proteome ChIP or "pChIP") in order to identify the different proteins interacting with RNAPII and components of the Polycomb complex; these experiments may uncover new RNAPII and Polycomb-associated proteins and thus provide insights into how their combinatorial binding may lead to diverse functional regulatory responses.

BIBLIOGRAPHY

S.H. Ahn, M. Kim, and S. Buratowski. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3'end processing. *Molecular cell*, 13(1):67–76, 2004.

M. Akhtar et al. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Molecular cell*, 34(3):387–393, 2009.

R.C. Akkers, S.J. van Heeringen, U.G. Jacobi, E.M. Janssen-Megens, K.J. Françoijs, H.G. Stunnenberg, and G.J.C. Veenstra. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Developmental cell*, 17 (3):425–434, 2009.

A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22 (13):1600, 2006.

S. Anders. Analysing RNA-Seq data with the DESeq package. *Molecular Biology*, pages 1–17, 2010.

R. Araki, R. Fukumura, N. Sasaki, Y. Kasama, N. Suzuki, H. Takahashi, Y. Tabata, T. Saito, and M. Abe. More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem Cells*, 24(11):2522–2528, 2006.

L. Atchison, A. Ghias, F. Wilkinson, N. Bonini, and M.L. Atchison. Transcription factor YY1 functions as a PcG protein in vivo. *The EMBO journal*, 22(6):1347–1358, 2003.

T. Atsuta, S. Fujimura, H. Moriya, M. Vidal, T. Akasaka, and H. Koseki. Production of monoclonal antibodies against mammalian Ring1B proteins. *Hybridoma*, 20(1): 43–46, 2001.

V. Azuara, P. Perry, S. Sauer, M. Spivakov, H.F. Jørgensen, R.M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, et al. Chromatin signatures of pluripotent cell lines. *Nature cell biology*, 8(5):532–538, 2006.

D. Baillat, M.A. Hakimi, A.M. Naar, A. Shilatifard, N. Cooch, and R. Shiekhattar. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, 123(2):265–276, 2005.

A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

R. Baskaran, SR Escobar, and JY Wang. Nuclear c-Abl is a COOH-terminal repeated domain (CTD)-tyrosine (CTD)-tyrosine kinase-specific for the mammalian RNA polymerase II: possible role in transcription elongation. *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, 10(6):387, 1999.

N.N. Batada and L.D. Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature genetics*, 39(8):945–950, 2007.

W. Bender and D.P. Fitzgerald. Transcription activates repressed domains in the Drosophila bithorax complex. *Development*, 129(21):4923, 2002.

B.E. Bernstein, T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.

N. Billon, C. Jolicoeur, Q.L. Ying, A. Smith, and M. Raff. Normal timing of oligodendrocyte development from genetically engineered, lineage-selectable mouse ES cells. *Journal of cell science*, 115(18):3657, 2002.

S. Bjorklund and C.M. Gustafsson. The yeast Mediator complex and its regulation. *Trends in biochemical sciences*, 30(5):240–244, 2005.

A.K. Boehm, A. Saunders, J. Werner, and J.T. Lis. Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. *Molecular and cellular biology*, 23(21):7628, 2003.

S. Boeing, C. Rigault, M. Heidemann, D. Eick, and M. Meisterernst. RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. *Journal of Biological Chemistry*, 285(1):188, 2010.

L.A. Boyer, T.I. Lee, M.F. Cole, S.E. Johnstone, S.S. Levine, J.P. Zucker, M.G. Guenther, R.M. Kumar, H.L. Murray, R.G. Jenner, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005.

L.A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L.A. Medeiros, T.I. Lee, S.S. Levine, M. Wernig, A. Tajonar, M.K. Ray, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–353, 2006.

M.R. Branco and A. Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, 4(5):e138, 2006.

D.B. Bregman, R.G. Pestell, and V.J. Kidd. Cell cycle regulation and RNA polymerase II. *Front Biosci*, 5(5):D244–257, 2000.

A. Breiling, L.P. O'Neill, D. D'Eliseo, B.M. Turner, and V. Orlando. Epigenome changes in active and inactive polycomb-group-controlled regions. *EMBO reports*, 5 (10):976–982, 2004.

A.S. Brodsky, C.A. Meyer, I.A. Swinburne, G. Hall, B.J. Keenan, X.S. Liu, E.A. Fox, and P.A. Silver. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome biology*, 6(8):R64, 2005.

E. Brookes. RNA Polymerase II regulation in Embryonic Stem cells, and its role in epigenetics and genome function. *PhD Thesis, Imperial College London*, 2010.

E. Brookes and A. Pombo. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO reports*, 10(11):1213–1219, 2009.

J.M. Brown, J. Green, R.P. das Neves, H.A.C. Wallace, A.J.H. Smith, J. Hughes, N. Gray, S. Taylor, W.G. Wood, D.R. Higgs, et al. Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *The Journal of Cell Biology*, 182(6):1083, 2008.

G. Buchwald, P. Van Der Stoop, O. Weichenrieder, A. Perrakis, M. Van Lohuizen, and T.K. Sixma. Structure and E3-ligase activity of the Ring-Ring complex of polycomb proteins Bmi1 and Ring1b. *The EMBO journal*, 25(11):2465–2474, 2006.

J. Bullard, E. Purdom, K. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

S. Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular cell*, 36(4):541–546, 2009.

M.A. Canham, A.A. Sharov, M.S.H. Ko, and J.M. Brickman. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biology*, 8(5):e1000379, 2010.

R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R.S. Jones, and Y. Zhang. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, 298(5595):1039, 2002.

R. Cao, Y. Tsukada, and Y. Zhang. Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Molecular cell*, 20(6):845–854, 2005.

P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P.G. Engstrom, M.C. Frith, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, 2006.

H. Cedar and Y. Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, 10(5):295–304, 2009.

I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234, 2007.

S. Chambeyron and W.A. Bickmore. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & development*, 18(10):1119, 2004.

R.D. Chapman, M. Conrad, and D. Eick. Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. *Molecular and cellular biology*, 25(17):7665, 2005.

R.D. Chapman, M. Heidemann, T.K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science*, 318(5857):1780, 2007.

R.D. Chapman, M. Heidemann, C. Hintermair, and D. Eick. Molecular evolution of the RNA polymerase II CTD. *Trends in Genetics*, 24(6):289–296, 2008.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.

B. Cheng and D.H. Price. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *Journal of Biological Chemistry*, 282(30):21901, 2007.

I. Chernukhin, S. Shamsuddin, S.Y. Kang, R. Bergstrom, Y.W. Kwon, W.Q. Yu, J. Whitehead, R. Mukhopadhyay, F. Docquier, D. Farrar, et al. CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Molecular and cellular biology*, 27(5):1631, 2007. ISSN 0270-7306.

E.J. Cho, M.S. Kobor, M. Kim, J. Greenblatt, and S. Buratowski. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. *Genes & development*, 15(24):3319, 2001.

V.S. Chopra, J. Cande, J.W. Hong, and M. Levine. Stalled Hox promoters as chromosomal boundaries. *Genes & development*, 23(13):1505, 2009a.

V.S. Chopra, J.W. Hong, and M. Levine. Regulation of Hox gene activity by transcriptional elongation in Drosophila. *Current biology*, 19(8):688–693, 2009b.

N. Cloonan, A.R.R. Forrest, G. Kolle, B.B.A. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–619, 2008.

R.C. Conaway, S. Sato, C. Tomomori-Sato, T. Yao, and J.W. Conaway. The mammalian Mediator complex and its role in transcriptional regulation. *Trends in biochemical sciences*, 30(5):250–255, 2005.

J.L. Corden, D.L. Cadena, J.M. Ahearn, and M.E. Dahmus. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proceedings of the National Academy of Sciences*, 82(23):7934, 1985.

L.J. Core, J.J. Waterfall, and J.T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845, 2008.

T. Cremer, M. Cremer, S. Dietzel, S. Muller, I. Solovei, and S. Fakan. Chromosome territories-a functional nuclear landscape. *Current opinion in cell biology*, 18(3): 307–316, 2006.

M.P. Creyghton, S. Markoulaki, S.S. Levine, J. Hanna, M.A. Lodato, K. Sha, R.A. Young, R. Jaenisch, and L.A. Boyer. H2az is enriched at polycomb complex target genes in es cells and is necessary for lineage commitment. *Cell*, 135(4):649–661, 2008.

K. Cui, C. Zang, T.Y. Roh, D.E. Schones, R.W. Childs, W. Peng, and K. Zhao. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, 4(1):80–93, 2009.

S. Dambacher, M. Hahn, and G. Schotta. Epigenetic regulation of development by histone lysine methylation. *Heredity*, 105(1):24–37, 2010.

S. De and M.M. Babu. Genomic neighbourhood and the regulation of gene expression. *Current Opinion in Cell Biology*, 22(3):326–333, 2010.

S. De, S.A. Teichmann, and M.M. Babu. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome research*, 19(5):785, 2009.

A.M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010, 2011.

J. Dekker. A closer look at long-range chromosomal interactions. *Trends in biochemical sciences*, 28(6):277–280, 2003.

M. del Mar Lorente, C. Marcos-Gutiérrez, C. Pérez, J. Schoorlemmer, A. Ramírez, T. Magin, and M. Vidal. Loss-and gain-of-function mutations show a polycomb group function for Ring1A in mice. *Development*, 127(23):5093, 2000.

G.I. Dellino, Y.B. Schwartz, G. Farkas, D. McCabe, S.C.R. Elgin, and V. Pirrotta. Polycomb silencing blocks transcription initiation. *Molecular cell*, 13(6):887–893, 2004.

G. Dieci, G. Fiorino, M. Castelnuovo, M. Teichmann, and A. Pagano. The expanding RNA polymerase III transcriptome. *Trends in Genetics*, 23(12):614–622, 2007.

H. Edelsbrunner and H.A. Maurer. On the intersection of orthogonal objects. *Information Processing Letters*, 13(4/5):177–181, 1981.

J.W. Edmunds, L.C. Mahadevan, and A.L. Clayton. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *The EMBO journal*, 27(2):406–420, 2007.

Y.J.K. Edwards, K. Bryson, and D.T. Jones. A meta-analysis of microarray gene expression in mouse stem cells: redefining stemness. *PLoS One*, 3(7):e2712, 2008.

S. Efroni, R. Duttagupta, J. Cheng, H. Dehghani, D.J. Hoeppner, C. Dash, D.P. Bazett-Jones, S. Le Grice, R.D.G. McKay, K.H. Buetow, et al. Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell*, 2(5):437–447, 2008.

S. Egloff, D. O'Reilly, R.D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science*, 318(5857):1777, 2007.

S. Egloff, S.A. Szczepaniak, M. Dienstbier, A. Taylor, S. Knight, and S. Murphy. The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *Journal of Biological Chemistry*, 285(27):20564, 2010.

M. Endoh, T.A. Endo, T. Endoh, Y. Fujimura, O. Ohara, T. Toyoda, A.P. Otte, M. Okano, N. Brockdorff, M. Vidal, et al. Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. *Development*, 135(8):1513, 2008.

R. Eskeland, M. Leeb, G.R. Grimes, C. Kress, S. Boyle, D. Sproul, N. Gilbert, Y. Fan, A.I. Skoultchi, A. Wutz, et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell*, 38(3): 452–464, 2010.

MJ Evans, MH Kaufman, et al. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, 1981.

C. Ferrai, S.Q. Xie, P. Luraghi, D. Munari, F. Ramirez, M.R. Branco, A. Pombo, and M.P. Crippa. Poised transcription factories prime silent upa gene prior to activation. *PLoS Biology*, 8(1):e1000270, 2010.

R. Festenstein, M. Tolaini, P. Corbella, C. Mamalaki, J. Parrington, M. Fox, A. Miliou, M. Jones, and D. Kioussis. Locus control region function and heterochromatin-induced position effect variegation. *Science*, 271(5252):1123, 1996.

P.C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of Drosophila and human core promoters. *Genome biology*, 7(7):R53, 2006.

J.F. Flanagan, L.Z. Mi, M. Chruszcz, M. Cymborowski, K.L. Clines, Y. Kim, W. Minor, F. Rastinejad, and S. Khorasanizadeh. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, 438(7071):1181–1185, 2005.

S.D. Fouse, Y. Shen, M. Pellegrini, S. Cole, A. Meissner, L. Van Neste, R. Jaenisch, and G. Fan. Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell*, 2(2):160–169, 2008.

N.J. Francis, R.E. Kingston, and C.L. Woodcock. Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574, 2004.

N.J. Fuda, M.B. Ardehali, and J.T. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192, 2009.

K. Fujinaga, D. Irwin, Y. Huang, R. Taube, T. Kurosu, and B.M. Peterlin. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Molecular and cellular biology*, 24(2):787, 2004.

M. Gardiner-Garden and M. Frommer. CpG Islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282, 1987.

A. Gaspar-Maia, A. Alajem, F. Polesso, R. Sridharan, M.J. Mason, A. Heidersbach, J. Ramalho-Santos, M.T. McManus, K. Plath, E. Meshorer, et al. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature*, 460(7257): 863–868, 2009.

A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451, 2005.

J.D. Gibson, C.M. Jakuba, N. Boucher, K.A. Holbrook, M.G. Carter, and C.E. Nelson. Single-cell transcript analysis of human embryonic stem cells. *Integr. Biol.*, 1(8-9): 540–551, 2009.

D.A. Gilchrist, S. Nechaev, C. Lee, S.K.B. Ghosh, J.B. Collins, L. Li, D.S. Gilmour, and K. Adelman. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes & Development*, 22(14): 1921, 2008.

D.A. Gilchrist, G. Dos Santos, D.C. Fargo, B. Xie, Y. Gao, L. Li, and K. Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551, 2010.

K. Glover-Cutter, S. Larochelle, B. Erickson, C. Zhang, K. Shokat, R.P. Fisher, and D.L. Bentley. TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. *Molecular and cellular biology*, 29(20):5455, 2009.

J.A. Goodrich and J.F. Kugel. Non-coding-RNA regulators of RNA polymerase II transcription. *Nature Reviews Molecular Cell Biology*, 7(8):612–616, 2006.

J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

T. Graf and M. Stadtfeld. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell*, 3(5):480–483, 2008.

J. Graumann, N.C. Hubner, J.B. Kim, K. Ko, M. Moser, C. Kumar, J. Cox, H. Scholer, and M. Mann. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Molecular & Cellular Proteomics*, 7(4):672, 2008.

I.H. Greger, A. Aranda, and N. Proudfoot. Balancing transcriptional interference and initiation on the GAL7 promoter of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 97(15):8415, 2000.

M.G. Guenther, S.S. Levine, L.A. Boyer, R. Jaenisch, and R.A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130 (1):77–88, 2007. ISSN 0092-8674.

R.A. Gupta, N. Shah, K.C. Wang, J. Kim, H.M. Horlings, D.J. Wong, M.C. Tsai, T. Hung, P. Argani, J.L. Rinn, et al. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071, 2010.

J.L. Gutierrez, M. Chandy, M.J. Carrozza, and J.L. Workman. Activation domains drive nucleosome eviction by swi/snf. *The EMBO Journal*, 26(3):730–740, 2007.

S.M. Hammond and N.E. Sharpless. HMGA2, microRNAs, and stem cell aging. *Cell*, 135(6):1013–1016, 2008.

J. Hani, B. Schelbert, A. Bernhardt, H. Domdey, G. Fischer, K. Wiebauer, and J.U. Rahfeld. Mutations in a Peptidylprolyl-cis/trans-isomerase Gene Lead to a Defect in 3-End Formation of a Pre-mRNA in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 274(1):108, 1999.

T. Hayakawa, Y. Ohtani, N. Hayakawa, K. Shinmyozu, M. Saito, F. Ishikawa, and J. Nakayama. RBP2 is an MRG15 complex component and down-regulates intragenic histone H3 lysine 4 methylation. *Genes to Cells*, 12(6):811–826, 2007.

K. Hayashi, S.M. Lopes, F. Tang, and M.A. Surani. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell stem cell*, 3(4):391–401, 2008.

D. Hebenstreit and S.A. Teichmann. Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Physical Biology*, 8:035013, 2011.

D. Hebenstreit, M. Gu, S. Haider, D.J. Turner, P. Liò, and S.A. Teichmann. EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic acids research*, 39(5):e27, 2011.

N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, L.O. Barrera, S. Van Calcar, C. Qu, K.A. Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007. ISSN 1061-4036.

C.J. Hengartner, V.E. Myer, S.M. Liao, C.J. Wilson, S.S. Koh, and R.A. Young. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Molecular cell*, 2(1):43–53, 1998.

S. Henikoff, T. Furuyama, and K. Ahmad. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics*, 20(7):320–326, 2004.

Y. Hirose and Y. Ohkuma. Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. *Journal of biochemistry*, 141(5):601, 2007.

C.K. Ho and S. Shuman. Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Molecular Cell*, 3(3):405–411, 1999.

I. Hogga and F. Karch. Transcription through the iab-7 cis-regulatory domain of the bithorax complex interferes with maintenance of Polycomb-mediated silencing. *Development*, 129(21):4915, 2002.

L.D. Hurst, C. Pal, and M.J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics*, 5(4):299–310, 2004.

P.J. Iaquinta and J.A. Lees. Life and death decisions by the E2F transcription factors. *Current opinion in cell biology*, 19(6):649–657, 2007.

F.J. Iborra, A. Pombo, D.A. Jackson, and P.R. Cook. Active RNA polymerases are localized within discrete transcription factories in human nuclei. *Journal of cell science*, 109(6):1427, 1996.

R.S. Illingworth and A.P. Bird. CpG islands-A rough guide. *FEBS letters*, 583(11): 1713–1720, 2009.

J.J.L. Jacobs, K. Kieboom, S. Marino, R.A. DePinho, and M. van Lohuizen. The oncogene and Polycomb-group gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature*, 397(6715):164–168, 1999.

S.C. Janga, J. Collado-Vides, and M.M. Babu. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761, 2008.

S.J. Jeong, H.J. Kim, Y.J. Yang, J.H. Seol, B.Y. Jung, J.W. Han, H.W. Lee, and E.J. Cho. Role of RNA polymerase II carboxy terminal domain phosphorylation in DNA damage response. *J Microbiol*, 43(6):516–522, 2005.

X. Ji, W. Li, J. Song, L. Wei, and X.S. Liu. CEAS: cis-regulatory element annotation system. *Nucleic acids research*, 34(suppl 2):W551, 2006.

T. Juven-Gershon and J.T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology*, 339(2): 225–229, 2010.

T. Juven-Gershon, J.Y. Hsu, J.W.M. Theisen, and J.T. Kadonaga. The RNA polymerase II core promoter–the gateway to transcription. *Current opinion in cell biology*, 20(3): 253–259, 2008.

J.T. Kadonaga. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 116(2):247–257, 2004.

M.H. Kagey, J.J. Newman, S. Bilodeau, Y. Zhan, D.A. Orlando, N.L. van Berkum, C.C. Ebmeier, J. Goossens, P.B. Rahl, S.S. Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27, 2000.

A. Kanhere, K. Viiri, C.C. Araujo, J. Rasaiyaah, R.D. Bouwman, W.A. Whyte, C.F. Pereira, E. Brookes, K. Walker, G.W. Bell, et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Molecular cell*, 38(5):675–688, 2010.

D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, YT Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, DJ Thomas, et al. The UCSC genome browser database. *Nucleic acids research*, 31(1):51, 2003.

B. Keenen and I.L. de la Serna. Chromatin remodeling in embryonic stem cells: regulating the balance between pluripotency and differentiation. *Journal of cellular physiology*, 219(1):1–7, 2009.

W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, et al. The human genome browser at UCSC. *Genome research*, 12(6):996, 2002.

M.C. Keogh, V. Podolny, and S. Buratowski. Bur1 kinase is required for efficient transcription elongation by RNA polymerase II. *Molecular and cellular biology*, 23 (19):7005, 2003.

A.M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B.E. Bernstein, A. Van Oudenaarden, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, 106(28):11667, 2009.

H. Kim, B. Erickson, W. Luo, D. Seward, J.H. Graber, D.D. Pollock, P.C. Megee, and D.L. Bentley. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nature structural & molecular biology*, 2010a.

J. Kim, J. Chu, X. Shen, J. Wang, and S.H. Orkin. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061, 2008.

T.H. Kim, L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond, Y. Wu, R.D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005. ISSN 0028-0836.

T.K. Kim, M. Hemberg, J.M. Gray, A.M. Costa, D.M. Bear, J. Wu, D.A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010b.

K.O. Kizer, H.P. Phatnani, Y. Shibata, H. Hall, A.L. Greenleaf, and B.D. Strahl. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Molecular and cellular biology*, 25(8): 3305, 2005.

F. Koch, F. Jourquin, P. Ferrier, and J.C. Andrau. Genome-wide RNA polymerase II: not genes only! *Trends in biochemical sciences*, 33(6):265–273, 2008.

P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X.S. Liu, and J. Ahringer. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, 41(3):376, 2009.

P. Komarnitsky, E.J. Cho, and S. Buratowski. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes & development*, 14(19):2452, 2000.

H. Kondoh, M.E. Lleonart, Y. Nakashima, M. Yokode, M. Tanaka, D. Bernard, J. Gil, and D. Beach. A high glycolytic flux supports the proliferative potential of murine embryonic stem cells. *Antioxidants & redox signaling*, 9(3):293–299, 2007.

A.R. Kornblihtt, M. de la Mata, J.P. Fededa, M.J. Munoz, and G. Nogues. Multiple links between transcription and splicing. *Rna*, 10(10):1489, 2004.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

S. Krishnamurthy, M.A. Ghazy, C. Moore, and M. Hampsey. Functional interaction of the Ess1 prolyl isomerase with components of the RNA polymerase II initiation and termination machineries. *Molecular and cellular biology*, 29(11):2925, 2009.

N.J. Krogan, M. Kim, A. Tong, A. Golshani, G. Cagney, V. Canadien, D.P. Richards, B.K. Beattie, A. Emili, C. Boone, et al. Methylation of histone H3 by Set2 in

Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Molecular and cellular biology*, 23(12):4207, 2003. ISSN 0270-7306.

M. Ku, R.P. Koche, E. Rheinbay, E.M. Mendenhall, M. Endoh, T.S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A.S. Chi, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics*, 4(10): e1000242, 2008. ISSN 1553-7404.

RM Kuhn, D. Karolchik, AS Zweig, T. Wang, KE Smith, KR Rosenbloom, B. Rhead, BJ Raney, A. Pohl, M. Pheasant, et al. The ucsc genome browser database: update 2009. *Nucleic acids research*, 37(suppl 1):D755, 2009.

D. Landeira and A.G. Fisher. Inactive yet indispensable: the tale of Jarid2. *Trends in cell biology*, pages 74–80, 2010.

D. Landeira, S. Sauer, R. Poot, M. Dvorkina, L. Mazzarella, H.F. Jørgensen, C.F. Pereira, M. Leleu, F.M. Piccolo, M. Spivakov, et al. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nature cell biology*, 12 (6):618–624, 2010.

B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

C. Lee, X. Li, A. Hechmer, M. Eisen, M.D. Biggin, B.J. Venters, C. Jiang, J. Li, B.F. Pugh, and D.S. Gilmour. NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Molecular and cellular biology*, 28(10):3290, 2008.

T.I. Lee and R.A. Young. Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, 34(1):77–137, 2000.

T.I. Lee, R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar, B. Chevalier, S.E. Johnstone, M.F. Cole, K. Isono, et al. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, 2006.

M. Leleu, G. Lefebvre, and J. Rougemont. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in functional genomics*, 9(5-6): 466, 2010.

M. Levine. Paused RNA Polymerase II as a Developmental Checkpoint. *Cell*, 145(4): 502–511, 2011.

B. Li, L.A. Howe, S. Anderson, J.R. Yates, and J.L. Workman. The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. *Journal of Biological Chemistry*, 278(11):8897, 2003.

B. Li, M. Carey, and J.L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007.

E. Li, T.H. Bestor, and R. Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 2009.

T. Li, J.F. Hu, X. Qiu, J. Ling, H. Chen, S. Wang, A. Hou, T.H. Vu, and A.R. Hoffman. CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. *Molecular and cellular biology*, 28(20): 6473, 2008.

S.M. Liao, J. Zhang, D.A. Jeffery, A.J. Koleske, C.M. Thompson, D.M. Chao, M. Viljoen, H.J.J. van Vuuren, and R.A. Young. A kinase-cyclin pair in the RNA polymerase II holoenzyme. *Nature*, 374(6518):193–196, 1995.

Y. Liu, C. Kung, J. Fishburn, A.Z. Ansari, K.M. Shokat, and S. Hahn. Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex. *Molecular and cellular biology*, 24(4):1721, 2004.

Y.H. Loh, Q. Wu, J.L. Chew, V.B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics*, 38(4):431–440, 2006.

X. Lu, M.D. Simon, J.V. Chodaparambil, J.C. Hansen, K.M. Shokat, and K. Luger. The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nature structural & molecular biology*, 15(10):1122, 2008.

R.F. Luco, Q. Pan, K. Tominaga, B.J. Blencowe, O.M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996, 2010.

A.H. Lund and M. van Lohuizen. Polycomb complexes and silencing mechanisms. *Current opinion in cell biology*, 16(3):239–246, 2004.

M. Madan Babu, S.C. Janga, I. de Santiago, and A. Pombo. Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current opinion in genetics & development*, 18(6):571–582, 2008.

Martin Maechler, Peter Rousseeuw, Anja Struyf, and Mia Hubert. Cluster Analysis Basics and Extensions. 2005.

J.R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nature genetics*, 38(10):1151–1158, 2006.

D. Mancini-DiNardo, S.J.S. Steele, R.S. Ingram, and S.M. Tilghman. A differentially methylated region within the gene kcnq1 functions as an imprinted promoter and silencer. *Human molecular genetics*, 12(3):283, 2003.

R. Margueron and D. Reinberg. The Polycomb complex PRC2 and its mark in life. *Nature*, 469(7330):343–349, 2011.

N.F. Marshall, J. Peng, Z. Xie, and D.H. Price. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *Journal of Biological Chemistry*, 271(43):27176, 1996.

G.R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634, 1981.

G.A. Maston, S.K. Evans, and M.R. Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

A. Mattout and E. Meshorer. Chromatin plasticity and genome organization in pluripotent embryonic stem cells. *Current Opinion in Cell Biology*, 22(3):334–341, 2010.

T. Max, M. Sogaard, and J.Q. Svejstrup. Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *Journal of Biological Chemistry*, 282(19):14113, 2007.

A. Mayer, M. Lidschreiber, M. Siebert, K. Leike, J. Soding, and P. Cramer. Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology*, 17(10):1272–1278, 2010.

K.J. Meaburn and T. Misteli. Chromosome territories. *Nature*, 445(7126):379–381, 2007.

A. Meinhart, T. Kamenski, S. Hoeppner, S. Baumli, and P. Cramer. A structural perspective of CTD function. *Genes & development*, 19(12):1401, 2005.

M. Meininghaus, R.D. Chapman, M. Horndasch, and D. Eick. Conditional expression of RNA polymerase II in mammalian cells. *Journal of Biological Chemistry*, 275(32): 24375, 2000.

A. Meissner, T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein, C. Nusbaum, D.B. Jaffe, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

E.M. Mendenhall, R.P. Koche, T. Truong, V.W. Zhou, B. Issac, A.S. Chi, M. Ku, and B.E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genetics*, 6:12, 2010.

E. Meshorer and T. Misteli. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature Reviews Molecular Cell Biology*, 7(7):540–546, 2006.

T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, et al. Genome-wide maps of chromatin state in

pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007. ISSN 0028-0836.

T.A. Milne, Y. Dou, M.E. Martin, H.W. Brock, R.G. Roeder, and J.L. Hess. MLL associates specifically with a subset of transcriptionally active target genes. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (41):14765, 2005.

I.M. Min, J.J. Waterfall, L.J. Core, R.J. Munroe, J. Schimenti, and J.T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes & Development*, 25(7):742, 2011.

A. Missra and D.S. Gilmour. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proceedings of the National Academy of Sciences*, 107(25):11301, 2010.

J.A. Mitchell and P. Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & development*, 22(1):20, 2008.

F. Mohn and D. Schubeler. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends in Genetics*, 25(3):129–136, 2009.

F. Mohn, M. Weber, M. Rebhan, T.C. Roloff, J. Richter, M.B. Stadler, M. Bibel, and D. Schubeler. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell*, 30(6): 755–766, 2008.

A.L. Mosley, S.G. Pattenden, M. Carey, S. Venkatesh, J.M. Gilmore, L. Florens, J.L. Workman, and M.P. Washburn. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Molecular cell*, 34(2):168–178, 2009.

K. Mousavi, H. Zare, A.H. Wang, and V. Sartorelli. Polycomb Protein Ezh1 Promotes RNA Polymerase II Elongation. *Molecular Cell*, 2011.

N.M. Mulholland, I.F.G. King, and R.E. Kingston. Regulation of Polycomb group complexes by the sequence-specific DNA binding proteins Zeste and GAGA. *Genes & development*, 17(22):2741, 2003.

G.W. Muse, D.A. Gilchrist, S. Nechaev, R. Shah, J.S. Parker, S.F. Grissom, J. Zeitlinger, and K. Adelman. RNA polymerase is poised for activation across the genome. *Nature genetics*, 39(12):1507–1511, 2007.

S. Nechaev, D.C. Fargo, G. Dos Santos, L. Liu, Y. Gao, and K. Adelman. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, 327(5963):335, 2010.

J.A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308, 1965.

J. Nevado, L. Gaudreau, M. Adam, and M. Ptashne. Transcriptional activation by artificial recruitment in mammalian cells. *Proceedings of the National Academy of Sciences*, 96(6):2674, 1999.

H.H. Ng, F. Robert, R.A. Young, and K. Struhl. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular cell*, 11(3):709–719, 2003.

H. Niwa, J. Miyazaki, and A.G. Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature genetics*, 24(4): 372–376, 2000.

M. Nonet, D. Sweetser, and R.A. Young. Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell*, 50(6):909–915, 1987.

J.E. Ohm, K.M. McGarvey, X. Yu, L. Cheng, K.E. Schuebel, L. Cope, H.P. Mohammad, W. Chen, V.C. Daniel, W. Yu, et al. A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics*, 39(2):237–242, 2007.

S.K.T. Ooi, C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S.P. Lin, C.D. Allis, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154):714–717, 2007.

C.S. Osborne, L. Chakalova, K.E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J.A. Mitchell, S. Lopes, W. Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10): 1065–1071, 2004.

A. Oshlack, M.D. Robinson, and M.D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.

Z. Ouyang, Q. Zhou, and W.H. Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521, 2009.

G. Pan, S. Tian, J. Nie, C. Yang, V. Ruotti, H. Wei, G.A. Jonsdottir, R. Stewart, and J.A. Thomson. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell stem cell*, 1(3):299–312, 2007.

R.R. Pandey, T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-DiNardo, and C. Kanduri. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell*, 32(2):232–246, 2008.

P.J. Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

Y. Pei, B. Schwer, and S. Shuman. Interactions between fission yeast Cdk9, its cyclin partner Pch1, and mRNA capping enzyme Pct1 suggest an elongation checkpoint for mRNA quality control. *Journal of Biological Chemistry*, 278(9):7180, 2003.

D. Peric-Hupkes, W. Meuleman, L. Pagie, S.W.M. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R.M. Kerkhoven, M. van Lohuizen, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4):603–613, 2010.

C.L. Peterson and J.L. Workman. Promoter targeting and chromatin remodeling by the swi/snf complex. *Current opinion in genetics & development*, 10(2):187–192, 2000.

K. Pfannkuche, H. Summer, O. Li, J. Hescheler, and P. Droge. The High Mobility Group Protein HMGA2: A Co-Regulator of Chromatin Structure and Pluripotency in Stem Cells? *Stem Cell Reviews and Reports*, 5(3):224–230, 2009.

N.J. Proudfoot, A. Furger, and M.J. Dye. Integrating mRNA processing with transcription. *Cell*, 108(4):501–512, 2002.

A.R. Quinlan and I.M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841, 2010.

R Development Core Team. R: A Language and Environment for Statistical Computing. 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

P.B. Rahl, C.Y. Lin, A.C. Seila, R.A. Flynn, S. McCuine, C.B. Burge, P.A. Sharp, and R.A. Young. c-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445, 2010. ISSN 0092-8674.

V.R. Ramirez-Carrozzi, D. Braas, D.M. Bhatt, C.S. Cheng, C. Hong, K.R. Doty, J.C. Black, A. Hoffmann, M. Carey, and S.T. Smale. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell*, 138(1):114–128, 2009.

A. Rechtsteiner, S. Ercan, T. Takasaki, T.M. Phippen, T.A. Egelhofer, W. Wang, H. Kimura, J.D. Lieb, and S. Strome. The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny. *PLoS Genetics*, 6(9):e1001091, 2010.

W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425, 2007.

P. Richard and J.L. Manley. Transcription termination by nuclear RNA polymerases. *Genes & development*, 23(11):1247, 2009.

J.E. Richardson. Fjoin: simple and efficient computation of feature overlaps. *Journal of Computational Biology*, 13(8):1457–1464, 2006.

H. Richly, M. Lange, E. Simboeck, and L. Di Croce. Setting and resetting of epigenetic marks in malignant transformation and development. *BioEssays*, 32(8):669–679, 2010.

P. Rickert, J.L. Corden, and E. Lees. Cyclin C/CDK8 and cyclin H/CDK7/p36 are biochemically distinct CTD kinases. *Oncogene*, 18(4):1093, 1999.

J.L. Rinn, M. Kertesz, J.K. Wang, S.L. Squazzo, X. Xu, S.A. Brugmann, L.H. Goodnough, J.A. Helms, P.J. Farnham, E. Segal, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, 2007.

C.R. Rodriguez, E.J. Cho, M.C. Keogh, C.L. Moore, A.L. Greenleaf, and S. Buratowski. Kin28, the TFIIH-associated carboxy-terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. *Molecular and Cellular Biology*, 20(1):104, 2000.

T.Y. Roh, S. Cuddapah, K. Cui, and K. Zhao. The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences*, 103 (43):15782, 2006.

A.E. Rougvie and J.T. Lis. The RNA polymerase II molecule at the 5'end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6): 795–804, 1988.

J.M. Rozenberg, A. Shlyakhtenko, K. Glass, V. Rishi, M.V. Myakishev, P.C. FitzGerald, and C. Vinson. All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC genomics*, 9(1): 67, 2008.

AI Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374, 2003.

A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D.A. Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6):424–436, 2007.

S. Schoeftner, A.K. Sengupta, S. Kubicek, K. Mechtler, L. Spahn, H. Koseki, T. Jenuwein, and A. Wutz. Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *The EMBO Journal*, 25(13): 3110–3122, 2006.

D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. Van Lohuizen, A. Tanay, and G. Cavalli. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS biology*, 7(1):e1000013, 2009.

Y.B. Schwartz and V. Pirrotta. Polycomb complexes and epigenetic states. *Current opinion in cell biology*, 20(3):266–273, 2008.

Y.B. Schwartz, T.G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta. Alternative epigenetic chromatin states of polycomb target genes. *PLoS genetics*, 6 (1):e1000805, 2010.

A.C. Seila, J.M. Calabrese, S.S. Levine, G.W. Yeo, P.B. Rahl, R.A. Flynn, R.A. Young, and P.A. Sharp. Divergent transcription from active promoters. *Science*, 322(5909): 1849, 2008.

K.H. Sene, C.J. Porter, G. Palidwor, C. Perez-Iratxeta, E.M. Muro, P.A. Campbell, M.A. Rudnicki, and M.A. Andrade-Navarro. Gene function in early mouse embryonic stem cell differentiation. *BMC genomics*, 8(1):85, 2007.

A.A. Sharov and M.S.H. Ko. Human ES cell profiling broadens the reach of bivalent domains. *Cell Stem Cell*, 1(3):237–238, 2007.

X. Shen, W. Kim, Y. Fujiwara, M.D. Simon, Y. Liu, M.R. Mysliwiec, G.C. Yuan, Y. Lee, and S.H. Orkin. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell*, 139(7):1303–1314, 2009.

J. Silva and A. Smith. Capturing pluripotency. *Cell*, 132(4):532–536, 2008.

M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature*, 200:6.

R.J. Sims, L.A. Rojas, D. Beck, R. Bonasio, R. Schuller, W.J. Drury, D. Eick, and D. Reinberg. The C-Terminal Domain of RNA Polymerase II Is Modified by Site-Specific Methylation. *Science*, 332(6025):99, 2011.

F. Sleutels, R. Zwart, and D.P. Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415(6873):810–813, 2002.

S.T. Smale and J.T. Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449, 2003.

M.V. Sobennikova, E.K. Shematorova, and G.V. Shpakovski. C-terminal domain of subunit Rpb1 of nuclear RNA polymerase II and its role in the transcription cycle. *Molecular Biology*, 41(3):387–401, 2007.

N. Spies, C.B. Nielsen, R.A. Padgett, and C.B. Burge. Biased chromatin signatures around polyadenylation sites and exons. *Molecular cell*, 36(2):245–254, 2009.

M. Spivakov and A.G. Fisher. Epigenetic signatures of stem-cell identity. *Nature Reviews Genetics*, 8(4):263–271, 2007.

D. Sproul, N. Gilbert, and W.A. Bickmore. The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics*, 6(10):775–781, 2005.

S.L. Squazzo, H. OGeen, V.M. Komashko, S.R. Krig, V.X. Jin, S. Jang, R. Margueron, D. Reinberg, R. Green, and P.J. Farnham. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome research*, 16(7):890, 2006.

A.L. Statham, D. Strbenac, M.W. Coolen, C. Stirzaker, S.J. Clark, and M.D. Robinson. Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics*, 26(13):1662, 2010.

E.J. Steinmetz, C.L. Warren, J.N. Kuehner, B. Panbehi, A.Z. Ansari, and D.A. Brow. Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Molecular cell*, 24(5):735–746, 2006. ISSN 1097-2765.

J.K. Stock, S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A.G. Fisher, and A. Pombo. Ring1-mediated ubiquitination of H2A restrains poised

RNA polymerase II at bivalent genes in mouse ES cells. *Nature cell biology*, 9(12): 1428–1435, 2007.

A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545, 2005.

H. Sutherland and W.A. Bickmore. Transcription factories: gene expression in unions? *Nature Reviews Genetics*, 10(7):457–466, 2009.

H. Szutorisz, N. Dillon, and L. Tora. The role of enhancers as centres for general transcription factor recruitment. *Trends in biochemical sciences*, 30(11):593–599, 2005.

P. Taghavi and M. van Lohuizen. Two paths to silence merge. *Nature*, 439:16, 2006.

M. Tahiliani, K.P. Koh, Y. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer, D.R. Liu, L. Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324 (5929):930, 2009.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

H. Takeshima, S. Yamashita, T. Shimazu, T. Niwa, and T. Ushijima. The presence of RNA polymerase II, active or stalled, predicts epigenetic fate of promoter CpG islands. *Genome research*, 19(11):1974, 2009.

A. Tanay, A.H. O'Donnell, M. Damelin, and T.H. Bestor. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proceedings of the National Academy of Sciences*, 104(13):5521, 2007.

N. Thakur, V.K. Tiwari, H. Thomassin, R.R. Pandey, M. Kanduri, A. Gondor, T. Grange, R. Ohlsson, and C. Kanduri. An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Molecular and cellular biology*, 24(18):7855, 2004.

J.P. Thomson, P.J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A.R.W. Kerr, A. Deaton, R. Andrews, K.D. James, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464(7291):1082–1086, 2010.

J.R. Tietjen, D.W. Zhang, J.B. Rodriguez-Molina, B.E. White, M.S. Akhtar, M. Heidemann, X. Li, R.D. Chapman, K. Shokat, S. Keles, et al. Chemical-genomic dissection of the CTD code. *Nature Structural & Molecular Biology*, 17(99):1154–1161, 2010.

H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcarcel, and R. Guigo. Nucleosome positioning as a determinant of exon recognition. *Nature structural & molecular biology*, 16(9):996–1001, 2009.

Y. Toyooka, D. Shimosato, K. Murakami, K. Takahashi, and H. Niwa. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development*, 135(5):909, 2008.

C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105, 2009.

C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

V. Tripathi, J.D. Ellis, Z. Shen, D.Y. Song, Q. Pan, A.T. Watt, S.M. Freier, C.F. Bennett, A. Sharma, P.A. Bubulya, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell*, 2010.

J. Trott, K. Hayashi, A. Surani, M.M. Babu, and A. Martinez-Arias. Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol. BioSyst.*, 2012.

B.M. Turner. Open chromatin and hypertranscription in embryonic stem cells. *Cell Stem Cell*, 2(5):408–410, 2008.

E. Valen, G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, M. Murata, H. Nishiyori, D. Lazarevic, D. Motti, et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome research*, 19(2):255, 2009.

P. Van Der Stoop, E.A. Boutsma, D. Hulsman, S. Noback, M. Heimerikx, R.M. Kerkhoven, J.W. Voncken, L.F.A. Wessels, and M. Van Lohuizen. Ubiquitin E3 ligase Ring1b/Rnf2 of polycomb repressive complex 1 contributes to stable maintenance of mouse embryonic stem cells. *PLoS One*, 3(5):e2235, 2008.

D. Van Hoof, R. Passier, D. Ward-Van Oostwaard, M.W.H. Pinkse, A.J.R. Heck, C.L. Mummery, and J. Krijgsveld. A quest for human and mouse embryonic stem cell-specific proteins. *Molecular & Cellular Proteomics*, 5(7):1261, 2006.

S. Varambally, S.M. Dhanasekaran, M. Zhou, T.R. Barrette, C. Kumar-Sinha, M.G. Sanda, D. Ghosh, K.J. Pienta, R. Sewalt, A.P. Otte, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, 419(6907): 624–629, 2002.

W.N. Venables and D.M. Smith. the R Development Core Team. *An Introduction to R, Version*, 1(0), 2003.

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304, 2001.

M. Vermeulen, K.W. Mulder, S. Denissov, WWM Pijnappel, F. van Schaik, R.A. Varier, M. Baltissen, H.G. Stunnenberg, M. Mann, H.T. Timmers, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131(1): 58–69, 2007.

R. Versteeg, B.D.C. Van Schaik, M.F. Van Batenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, and A.H.C. van Kampen. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome research*, 13(9):1998, 2003.

E. Vire, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J.M. Vanderwinden, et al. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439(7078):871–874, 2006.

J.W. Voncken, B.A.J. Roelen, M. Roefs, S. De Vries, E. Verhoeven, S. Marino, J. Deschamps, and M. Van Lohuizen. Rnf2 (Ring1b) deficiency causes gastrulation arrest and cell cycle inhibition. *Proceedings of the National Academy of Sciences*, 100 (5):2468, 2003.

HB Wang, L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R.S. Jones, and Y. Zhang. Role of H2A ubiquitination in Polycomb silencing. *Nature*, 431:873–878, 2004.

J. Wang, S. Rao, J. Chu, X. Shen, D.N. Levasseur, T.W. Theunissen, and S.H. Orkin. A protein interaction network for pluripotency of embryonic stem cells. *Tandem*, 1 (10):30, 2006.

L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1): 136, 2010.

M. Weber, I. Hellmann, M.B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature genetics*, 39(4):457–466, 2007.

G. Wei, L. Wei, J. Zhu, C. Zang, J. Hu-Li, Z. Yao, K. Cui, Y. Kanno, T.Y. Roh, W.T. Watford, et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*, 30(1):155–167, 2009.

M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D.J. Weisenberger, M. Campan, J. Young, I. Jacobs, et al. Epigenetic stem cell signature in cancer. *Nature genetics*, 39(2):157–158, 2006.

R.L. Williams, D.J. Hilton, S. Pease, T.A. Willson, C.L. Stewart, D.P. Gearing, E.F. Wagner, D. Metcalf, N.A. Nicola, and N.M. Gough. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature*, 336 (6200):684–687, 1988.

R.R.E. Williams, V. Azuara, P. Perry, S. Sauer, M. Dvorkina, H. Jørgensen, J. Roix, P. McQueen, T. Misteli, M. Merkenschlager, et al. Neural induction promotes

large-scale chromatin reorganisation of the Mash1 locus. *Journal of Cell Science*, 119 (1):132, 2006.

A. Wood and A. Shilatifard. Bur1, Bur2 and the Ctk complex in yeast: the split personality of mammalian P-TEFb. *Cell Cycle*, 5(10):1066–1068, 2006.

C.H. Wu, Y. Yamaguchi, L.R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes & development*, 17(11):1402, 2003.

H. Wu, A.C. D Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y.E. Sun, and Y. Zhang. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473(7347):389–393, 2011.

Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

J. Wysocka, T. Swigut, H. Xiao, T.A. Milne, S.Y. Kwon, J. Landry, M. Kauer, A.J. Tackett, B.T. Chait, P. Badenhorst, et al. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86, 2006.

T. Xiao, H. Hall, K.O. Kizer, Y. Shibata, M.C. Hall, C.H. Borchers, and B.D. Strahl. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. *Genes & Development*, 17(5):654, 2003. ISSN 0890-9369.

S.Q. Xie, S. Martin, P.V. Guillot, D.L. Bentley, and A. Pombo. Splicing speckles are not reservoirs of RNA polymerase II, but contain an inactive form, phosphorylated on serine2 residues of the C-terminal domain. *Molecular biology of the cell*, 17(4):1723, 2006.

H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C.L. Wei, F. Lin, and W.K. Sung. A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199, 2010.

Y.X. Xu and J.L. Manley. Pin1 modulates RNA polymerase II activity during the transcription cycle. *Genes & development*, 21(22):2950, 2007.

T. Yamada, Y. Yamaguchi, N. Inukai, S. Okamoto, T. Mura, and H. Handa. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Molecular cell*, 21(2):227–237, 2006.

Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51, 1999.

G.S. Yochum and D.E. Ayer. Role for the mortality factors MORF4, MRGX, and MRG15 in transcriptional repression via associations with Pf1, mSin3A, and Transducin-Like Enhancer of Split. *Molecular and cellular biology*, 22(22):7868, 2002.

M.D. Young, T.A. Willson, M.J. Wakefield, E. Trounson, D.J. Hilton, M.E. Blewitt, A. Oshlack, and I.J. Majewski. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Research*, 2011.

C. Zang, D.E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25(15):1952, 2009.

J. Zeitlinger, A. Stark, M. Kellis, J.W. Hong, S. Nechaev, K. Adelman, M. Levine, and R.A. Young. RNA Polymerase Stalling at Developmental Control Genes in the Drosophila Embryo. *Nature genetics*, 39(12):1512, 2007.

P. Zhang, J. Du, B. Sun, X. Dong, G. Xu, J. Zhou, Q. Huang, Q. Liu, Q. Hao, and J. Ding. Structure of human MRG15 chromo domain and its binding to Lys36-methylated histone H3. *Nucleic acids research*, 34(22):6621, 2006a.

Y. Zhang, Y. Kim, N. Genoud, J. Gao, J.W. Kelly, S.L. Pfaff, G.N. Gill, J.E. Dixon, and J.P. Noel. Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Molecular cell*, 24(5):759–770, 2006b.

J. Zhao, B.K. Sun, J.A. Erwin, J.J. Song, and J.T. Lee. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322(5902):750, 2008.

X.D. Zhao, X. Han, J.L. Chew, J. Liu, K.P. Chiu, A. Choo, Y.L. Orlov, W.K. Sung, A. Shahab, V.A. Kuznetsov, et al. Whole-genome mapping of histone H3 Lys4 and

27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell stem cell*, 1(3):286–298, 2007.

W. Zhou, P. Zhu, J. Wang, G. Pascual, K.A. Ohgi, J. Lozach, C.K. Glass, and M.G. Rosenfeld. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Molecular cell*, 29(1):69–80, 2008.

D. Zilberman, D. Coleman-Derr, T. Ballinger, and S. Henikoff. Histone H2A. Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218): 125, 2008.

Part V

APPENDIX

BIBLIOGRAPHY

S.H. Ahn, M. Kim, and S. Buratowski. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3'end processing. *Molecular cell*, 13(1):67–76, 2004.

M. Akhtar et al. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Molecular cell*, 34(3):387–393, 2009.

R.C. Akkers, S.J. van Heeringen, U.G. Jacobi, E.M. Janssen-Megens, K.J. Françoijs, H.G. Stunnenberg, and G.J.C. Veenstra. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Developmental cell*, 17 (3):425–434, 2009.

A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22 (13):1600, 2006.

S. Anders. Analysing RNA-Seq data with the DESeq package. *Molecular Biology*, pages 1–17, 2010.

R. Araki, R. Fukumura, N. Sasaki, Y. Kasama, N. Suzuki, H. Takahashi, Y. Tabata, T. Saito, and M. Abe. More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem Cells*, 24(11):2522–2528, 2006.

L. Atchison, A. Ghias, F. Wilkinson, N. Bonini, and M.L. Atchison. Transcription factor YY1 functions as a PcG protein in vivo. *The EMBO journal*, 22(6):1347–1358, 2003.

T. Atsuta, S. Fujimura, H. Moriya, M. Vidal, T. Akasaka, and H. Koseki. Production of monoclonal antibodies against mammalian Ring1B proteins. *Hybridoma*, 20(1): 43–46, 2001.

V. Azuara, P. Perry, S. Sauer, M. Spivakov, H.F. Jørgensen, R.M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, et al. Chromatin signatures of pluripotent cell lines. *Nature cell biology*, 8(5):532–538, 2006.

D. Baillat, M.A. Hakimi, A.M. Naar, A. Shilatifard, N. Cooch, and R. Shiekhattar. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, 123(2):265–276, 2005.

A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

R. Baskaran, SR Escobar, and JY Wang. Nuclear c-Abl is a COOH-terminal repeated domain (CTD)-tyrosine (CTD)-tyrosine kinase-specific for the mammalian RNA polymerase II: possible role in transcription elongation. *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, 10(6):387, 1999.

N.N. Batada and L.D. Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature genetics*, 39(8):945–950, 2007.

W. Bender and D.P. Fitzgerald. Transcription activates repressed domains in the Drosophila bithorax complex. *Development*, 129(21):4923, 2002.

B.E. Bernstein, T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.

N. Billon, C. Jolicoeur, Q.L. Ying, A. Smith, and M. Raff. Normal timing of oligodendrocyte development from genetically engineered, lineage-selectable mouse ES cells. *Journal of cell science*, 115(18):3657, 2002.

S. Bjorklund and C.M. Gustafsson. The yeast Mediator complex and its regulation. *Trends in biochemical sciences*, 30(5):240–244, 2005.

A.K. Boehm, A. Saunders, J. Werner, and J.T. Lis. Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. *Molecular and cellular biology*, 23(21):7628, 2003.

S. Boeing, C. Rigault, M. Heidemann, D. Eick, and M. Meisterernst. RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. *Journal of Biological Chemistry*, 285(1):188, 2010.

L.A. Boyer, T.I. Lee, M.F. Cole, S.E. Johnstone, S.S. Levine, J.P. Zucker, M.G. Guenther, R.M. Kumar, H.L. Murray, R.G. Jenner, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005.

L.A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L.A. Medeiros, T.I. Lee, S.S. Levine, M. Wernig, A. Tajonar, M.K. Ray, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–353, 2006.

M.R. Branco and A. Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, 4(5):e138, 2006.

D.B. Bregman, R.G. Pestell, and V.J. Kidd. Cell cycle regulation and RNA polymerase II. *Front Biosci*, 5(5):D244–257, 2000.

A. Breiling, L.P. O'Neill, D. D'Eliseo, B.M. Turner, and V. Orlando. Epigenome changes in active and inactive polycomb-group-controlled regions. *EMBO reports*, 5(10): 976–982, 2004.

A.S. Brodsky, C.A. Meyer, I.A. Swinburne, G. Hall, B.J. Keenan, X.S. Liu, E.A. Fox, and P.A. Silver. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome biology*, 6(8):R64, 2005.

E. Brookes. RNA Polymerase II regulation in Embryonic Stem cells, and its role in epigenetics and genome function. *PhD Thesis, Imperial College London*, 2010.

E. Brookes and A. Pombo. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO reports*, 10(11):1213–1219, 2009.

J.M. Brown, J. Green, R.P. das Neves, H.A.C. Wallace, A.J.H. Smith, J. Hughes, N. Gray, S. Taylor, W.G. Wood, D.R. Higgs, et al. Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *The Journal of Cell Biology*, 182(6):1083, 2008.

G. Buchwald, P. Van Der Stoop, O. Weichenrieder, A. Perrakis, M. Van Lohuizen, and T.K. Sixma. Structure and E3-ligase activity of the Ring-Ring complex of polycomb proteins Bmi1 and Ring1b. *The EMBO journal*, 25(11):2465–2474, 2006.

J. Bullard, E. Purdom, K. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

S. Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular cell*, 36(4):541–546, 2009.

M.A. Canham, A.A. Sharov, M.S.H. Ko, and J.M. Brickman. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biology*, 8(5):e1000379, 2010.

R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R.S. Jones, and Y. Zhang. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, 298(5595):1039, 2002.

R. Cao, Y. Tsukada, and Y. Zhang. Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Molecular cell*, 20(6):845–854, 2005.

P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P.G. Engstrom, M.C. Frith, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, 2006.

H. Cedar and Y. Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, 10(5):295–304, 2009.

I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234, 2007.

S. Chambeyron and W.A. Bickmore. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & development*, 18(10):1119, 2004.

R.D. Chapman, M. Conrad, and D. Eick. Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. *Molecular and cellular biology*, 25(17):7665, 2005.

R.D. Chapman, M. Heidemann, T.K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science*, 318(5857):1780, 2007.

R.D. Chapman, M. Heidemann, C. Hintermair, and D. Eick. Molecular evolution of the RNA polymerase II CTD. *Trends in Genetics*, 24(6):289–296, 2008.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.

B. Cheng and D.H. Price. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *Journal of Biological Chemistry*, 282(30):21901, 2007.

I. Chernukhin, S. Shamsuddin, S.Y. Kang, R. Bergstrom, Y.W. Kwon, W.Q. Yu, J. Whitehead, R. Mukhopadhyay, F. Docquier, D. Farrar, et al. CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Molecular and cellular biology*, 27(5):1631, 2007. ISSN 0270-7306.

E.J. Cho, M.S. Kobor, M. Kim, J. Greenblatt, and S. Buratowski. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. *Genes & development*, 15(24):3319, 2001.

V.S. Chopra, J. Cande, J.W. Hong, and M. Levine. Stalled Hox promoters as chromosomal boundaries. *Genes & development*, 23(13):1505, 2009a.

V.S. Chopra, J.W. Hong, and M. Levine. Regulation of Hox gene activity by transcriptional elongation in Drosophila. *Current biology*, 19(8):688–693, 2009b.

N. Cloonan, A.R.R. Forrest, G. Kolle, B.B.A. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–619, 2008.

R.C. Conaway, S. Sato, C. Tomomori-Sato, T. Yao, and J.W. Conaway. The mammalian Mediator complex and its role in transcriptional regulation. *Trends in biochemical sciences*, 30(5):250–255, 2005.

J.L. Corden, D.L. Cadena, J.M. Ahearn, and M.E. Dahmus. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proceedings of the National Academy of Sciences*, 82(23):7934, 1985.

L.J. Core, J.J. Waterfall, and J.T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845, 2008.

T. Cremer, M. Cremer, S. Dietzel, S. Muller, I. Solovei, and S. Fakan. Chromosome territories-a functional nuclear landscape. *Current opinion in cell biology*, 18(3):307–316, 2006.

M.P. Creyghton, S. Markoulaki, S.S. Levine, J. Hanna, M.A. Lodato, K. Sha, R.A. Young, R. Jaenisch, and L.A. Boyer. H2az is enriched at polycomb complex target genes in es cells and is necessary for lineage commitment. *Cell*, 135(4):649–661, 2008.

K. Cui, C. Zang, T.Y. Roh, D.E. Schones, R.W. Childs, W. Peng, and K. Zhao. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, 4(1):80–93, 2009.

S. Dambacher, M. Hahn, and G. Schotta. Epigenetic regulation of development by histone lysine methylation. *Heredity*, 105(1):24–37, 2010.

S. De and M.M. Babu. Genomic neighbourhood and the regulation of gene expression. *Current Opinion in Cell Biology*, 22(3):326–333, 2010.

S. De, S.A. Teichmann, and M.M. Babu. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome research*, 19(5):785, 2009.

A.M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010, 2011.

J. Dekker. A closer look at long-range chromosomal interactions. *Trends in biochemical sciences*, 28(6):277–280, 2003.

M. del Mar Lorente, C. Marcos-Gutiérrez, C. Pérez, J. Schoorlemmer, A. Ramírez, T. Magin, and M. Vidal. Loss-and gain-of-function mutations show a polycomb group function for Ring1A in mice. *Development*, 127(23):5093, 2000.

G.I. Dellino, Y.B. Schwartz, G. Farkas, D. McCabe, S.C.R. Elgin, and V. Pirrotta. Polycomb silencing blocks transcription initiation. *Molecular cell*, 13(6):887–893, 2004.

G. Dieci, G. Fiorino, M. Castelnuovo, M. Teichmann, and A. Pagano. The expanding RNA polymerase III transcriptome. *Trends in Genetics*, 23(12):614–622, 2007.

H. Edelsbrunner and H.A. Maurer. On the intersection of orthogonal objects. *Information Processing Letters*, 13(4/5):177–181, 1981.

J.W. Edmunds, L.C. Mahadevan, and A.L. Clayton. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *The EMBO journal*, 27(2):406–420, 2007.

Y.J.K. Edwards, K. Bryson, and D.T. Jones. A meta-analysis of microarray gene expression in mouse stem cells: redefining stemness. *PLoS One*, 3(7):e2712, 2008.

S. Efroni, R. Duttagupta, J. Cheng, H. Dehghani, D.J. Hoeppner, C. Dash, D.P. Bazett-Jones, S. Le Grice, R.D.G. McKay, K.H. Buetow, et al. Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell*, 2(5):437–447, 2008.

S. Egloff, D. O'Reilly, R.D. Chapman, A. Taylor, K. Tanzhaus, L. Pitts, D. Eick, and S. Murphy. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science*, 318(5857):1777, 2007.

S. Egloff, S.A. Szczepaniak, M. Dienstbier, A. Taylor, S. Knight, and S. Murphy. The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *Journal of Biological Chemistry*, 285(27):20564, 2010.

M. Endoh, T.A. Endo, T. Endoh, Y. Fujimura, O. Ohara, T. Toyoda, A.P. Otte, M. Okano, N. Brockdorff, M. Vidal, et al. Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. *Development*, 135(8):1513, 2008.

R. Eskeland, M. Leeb, G.R. Grimes, C. Kress, S. Boyle, D. Sproul, N. Gilbert, Y. Fan, A.I. Skoultchi, A. Wutz, et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell*, 38(3):452–464, 2010.

MJ Evans, MH Kaufman, et al. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, 1981.

C. Ferrai, S.Q. Xie, P. Luraghi, D. Munari, F. Ramirez, M.R. Branco, A. Pombo, and M.P. Crippa. Poised transcription factories prime silent upa gene prior to activation. *PLoS Biology*, 8(1):e1000270, 2010.

R. Festenstein, M. Tolaini, P. Corbella, C. Mamalaki, J. Parrington, M. Fox, A. Miliou, M. Jones, and D. Kioussis. Locus control region function and heterochromatin-induced position effect variegation. *Science*, 271(5252):1123, 1996.

P.C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of Drosophila and human core promoters. *Genome biology*, 7(7):R53, 2006.

J.F. Flanagan, L.Z. Mi, M. Chruszcz, M. Cymborowski, K.L. Clines, Y. Kim, W. Minor, F. Rastinejad, and S. Khorasanizadeh. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, 438(7071):1181–1185, 2005.

S.D. Fouse, Y. Shen, M. Pellegrini, S. Cole, A. Meissner, L. Van Neste, R. Jaenisch, and G. Fan. Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell*, 2(2):160–169, 2008.

N.J. Francis, R.E. Kingston, and C.L. Woodcock. Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574, 2004.

N.J. Fuda, M.B. Ardehali, and J.T. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192, 2009.

K. Fujinaga, D. Irwin, Y. Huang, R. Taube, T. Kurosu, and B.M. Peterlin. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Molecular and cellular biology*, 24(2):787, 2004.

M. Gardiner-Garden and M. Frommer. CpG Islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282, 1987.

A. Gaspar-Maia, A. Alajem, F. Polesso, R. Sridharan, M.J. Mason, A. Heidersbach, J. Ramalho-Santos, M.T. McManus, K. Plath, E. Meshorer, et al. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature*, 460(7257):863–868, 2009.

A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451, 2005.

J.D. Gibson, C.M. Jakuba, N. Boucher, K.A. Holbrook, M.G. Carter, and C.E. Nelson. Single-cell transcript analysis of human embryonic stem cells. *Integr. Biol.*, 1(8-9): 540–551, 2009.

D.A. Gilchrist, S. Nechaev, C. Lee, S.K.B. Ghosh, J.B. Collins, L. Li, D.S. Gilmour, and K. Adelman. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes & Development*, 22(14): 1921, 2008.

D.A. Gilchrist, G. Dos Santos, D.C. Fargo, B. Xie, Y. Gao, L. Li, and K. Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551, 2010.

K. Glover-Cutter, S. Larochelle, B. Erickson, C. Zhang, K. Shokat, R.P. Fisher, and D.L. Bentley. TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. *Molecular and cellular biology*, 29(20):5455, 2009.

J.A. Goodrich and J.F. Kugel. Non-coding-RNA regulators of RNA polymerase II transcription. *Nature Reviews Molecular Cell Biology*, 7(8):612–616, 2006.

J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

T. Graf and M. Stadtfeld. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell*, 3(5):480–483, 2008.

J. Graumann, N.C. Hubner, J.B. Kim, K. Ko, M. Moser, C. Kumar, J. Cox, H. Scholer, and M. Mann. Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Molecular & Cellular Proteomics*, 7(4):672, 2008.

I.H. Greger, A. Aranda, and N. Proudfoot. Balancing transcriptional interference and initiation on the GAL7 promoter of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 97(15):8415, 2000.

M.G. Guenther, S.S. Levine, L.A. Boyer, R. Jaenisch, and R.A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130 (1):77–88, 2007. ISSN 0092-8674.

R.A. Gupta, N. Shah, K.C. Wang, J. Kim, H.M. Horlings, D.J. Wong, M.C. Tsai, T. Hung, P. Argani, J.L. Rinn, et al. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071, 2010.

J.L. Gutierrez, M. Chandy, M.J. Carrozza, and J.L. Workman. Activation domains drive nucleosome eviction by swi/snf. *The EMBO Journal*, 26(3):730–740, 2007.

S.M. Hammond and N.E. Sharpless. HMGA2, microRNAs, and stem cell aging. *Cell*, 135(6):1013–1016, 2008.

J. Hani, B. Schelbert, A. Bernhardt, H. Domdey, G. Fischer, K. Wiebauer, and J.U. Rahfeld. Mutations in a Peptidylprolyl-cis/trans-isomerase Gene Lead to a Defect in 3-End Formation of a Pre-mRNA in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 274(1):108, 1999.

T. Hayakawa, Y. Ohtani, N. Hayakawa, K. Shinmyozu, M. Saito, F. Ishikawa, and J. Nakayama. RBP2 is an MRG15 complex component and down-regulates intragenic histone H3 lysine 4 methylation. *Genes to Cells*, 12(6):811–826, 2007.

K. Hayashi, S.M. Lopes, F. Tang, and M.A. Surani. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell stem cell*, 3(4):391–401, 2008.

D. Hebenstreit and S.A. Teichmann. Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Physical Biology*, 8:035013, 2011.

D. Hebenstreit, M. Gu, S. Haider, D.J. Turner, P. Liò, and S.A. Teichmann. EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic acids research*, 39(5):e27, 2011.

N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, L.O. Barrera, S. Van Calcar, C. Qu, K.A. Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007. ISSN 1061-4036.

C.J. Hengartner, V.E. Myer, S.M. Liao, C.J. Wilson, S.S. Koh, and R.A. Young. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Molecular cell*, 2(1):43–53, 1998.

S. Henikoff, T. Furuyama, and K. Ahmad. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics*, 20(7):320–326, 2004.

Y. Hirose and Y. Ohkuma. Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. *Journal of biochemistry*, 141(5):601, 2007.

C.K. Ho and S. Shuman. Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Molecular Cell*, 3(3):405–411, 1999.

I. Hogga and F. Karch. Transcription through the iab-7 cis-regulatory domain of the bithorax complex interferes with maintenance of Polycomb-mediated silencing. *Development*, 129(21):4915, 2002.

L.D. Hurst, C. Pal, and M.J. Lercher. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics*, 5(4):299–310, 2004.

P.J. Iaquinta and J.A. Lees. Life and death decisions by the E2F transcription factors. *Current opinion in cell biology*, 19(6):649–657, 2007.

F.J. Iborra, A. Pombo, D.A. Jackson, and P.R. Cook. Active RNA polymerases are localized within discrete transcription factories in human nuclei. *Journal of cell science*, 109(6):1427, 1996.

R.S. Illingworth and A.P. Bird. CpG islands-A rough guide. *FEBS letters*, 583(11): 1713–1720, 2009.

J.J.L. Jacobs, K. Kieboom, S. Marino, R.A. DePinho, and M. van Lohuizen. The oncogene and Polycomb-group gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature*, 397(6715):164–168, 1999.

S.C. Janga, J. Collado-Vides, and M.M. Babu. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761, 2008.

S.J. Jeong, H.J. Kim, Y.J. Yang, J.H. Seol, B.Y. Jung, J.W. Han, H.W. Lee, and E.J. Cho. Role of RNA polymerase II carboxy terminal domain phosphorylation in DNA damage response. *J Microbiol*, 43(6):516–522, 2005.

X. Ji, W. Li, J. Song, L. Wei, and X.S. Liu. CEAS: cis-regulatory element annotation system. *Nucleic acids research*, 34(suppl 2):W551, 2006.

T. Juven-Gershon and J.T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology*, 339(2): 225–229, 2010.

T. Juven-Gershon, J.Y. Hsu, J.W.M. Theisen, and J.T. Kadonaga. The RNA polymerase II core promoter–the gateway to transcription. *Current opinion in cell biology*, 20(3): 253–259, 2008.

J.T. Kadonaga. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 116(2):247–257, 2004.

M.H. Kagey, J.J. Newman, S. Bilodeau, Y. Zhan, D.A. Orlando, N.L. van Berkum, C.C. Ebmeier, J. Goossens, P.B. Rahl, S.S. Levine, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27, 2000.

A. Kanhere, K. Viiri, C.C. Araujo, J. Rasaiyaah, R.D. Bouwman, W.A. Whyte, C.F. Pereira, E. Brookes, K. Walker, G.W. Bell, et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Molecular cell*, 38(5):675–688, 2010.

D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, YT Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, DJ Thomas, et al. The UCSC genome browser database. *Nucleic acids research*, 31(1):51, 2003.

B. Keenen and I.L. de la Serna. Chromatin remodeling in embryonic stem cells: regulating the balance between pluripotency and differentiation. *Journal of cellular physiology*, 219(1):1–7, 2009.

W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, et al. The human genome browser at UCSC. *Genome research*, 12(6):996, 2002.

M.C. Keogh, V. Podolny, and S. Buratowski. Bur1 kinase is required for efficient transcription elongation by RNA polymerase II. *Molecular and cellular biology*, 23 (19):7005, 2003.

A.M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B.E. Bernstein, A. Van Oudenaarden, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, 106(28):11667, 2009.

H. Kim, B. Erickson, W. Luo, D. Seward, J.H. Graber, D.D. Pollock, P.C. Megee, and D.L. Bentley. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nature structural & molecular biology*, 2010a.

J. Kim, J. Chu, X. Shen, J. Wang, and S.H. Orkin. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061, 2008.

T.H. Kim, L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond, Y. Wu, R.D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005. ISSN 0028-0836.

T.K. Kim, M. Hemberg, J.M. Gray, A.M. Costa, D.M. Bear, J. Wu, D.A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010b.

K.O. Kizer, H.P. Phatnani, Y. Shibata, H. Hall, A.L. Greenleaf, and B.D. Strahl. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Molecular and cellular biology*, 25(8): 3305, 2005.

F. Koch, F. Jourquin, P. Ferrier, and J.C. Andrau. Genome-wide RNA polymerase II: not genes only! *Trends in biochemical sciences*, 33(6):265–273, 2008.

P. Kolasinska-Zwierz, T. Down, I. Latorre, T. Liu, X.S. Liu, and J. Ahringer. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics*, 41(3):376, 2009.

P. Komarnitsky, E.J. Cho, and S. Buratowski. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes & development*, 14(19):2452, 2000.

H. Kondoh, M.E. Lleonart, Y. Nakashima, M. Yokode, M. Tanaka, D. Bernard, J. Gil, and D. Beach. A high glycolytic flux supports the proliferative potential of murine embryonic stem cells. *Antioxidants & redox signaling*, 9(3):293–299, 2007.

A.R. Kornblihtt, M. de la Mata, J.P. Fededa, M.J. Munoz, and G. Nogues. Multiple links between transcription and splicing. *Rna*, 10(10):1489, 2004.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

S. Krishnamurthy, M.A. Ghazy, C. Moore, and M. Hampsey. Functional interaction of the Ess1 prolyl isomerase with components of the RNA polymerase II initiation and termination machineries. *Molecular and cellular biology*, 29(11):2925, 2009.

N.J. Krogan, M. Kim, A. Tong, A. Golshani, G. Cagney, V. Canadien, D.P. Richards, B.K. Beattie, A. Emili, C. Boone, et al. Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Molecular and cellular biology*, 23(12):4207, 2003. ISSN 0270-7306.

M. Ku, R.P. Koche, E. Rheinbay, E.M. Mendenhall, M. Endoh, T.S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A.S. Chi, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics*, 4(10): e1000242, 2008. ISSN 1553-7404.

RM Kuhn, D. Karolchik, AS Zweig, T. Wang, KE Smith, KR Rosenbloom, B. Rhead, BJ Raney, A. Pohl, M. Pheasant, et al. The ucsc genome browser database: update 2009. *Nucleic acids research*, 37(suppl 1):D755, 2009.

D. Landeira and A.G. Fisher. Inactive yet indispensable: the tale of Jarid2. *Trends in cell biology*, pages 74–80, 2010.

D. Landeira, S. Sauer, R. Poot, M. Dvorkina, L. Mazzarella, H.F. Jørgensen, C.F. Pereira, M. Leleu, F.M. Piccolo, M. Spivakov, et al. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nature cell biology*, 12(6):618–624, 2010.

B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

C. Lee, X. Li, A. Hechmer, M. Eisen, M.D. Biggin, B.J. Venters, C. Jiang, J. Li, B.F. Pugh, and D.S. Gilmour. NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Molecular and cellular biology*, 28(10):3290, 2008.

T.I. Lee and R.A. Young. Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, 34(1):77–137, 2000.

T.I. Lee, R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar, B. Chevalier, S.E. Johnstone, M.F. Cole, K. Isono, et al. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, 125(2):301–313, 2006.

M. Leleu, G. Lefebvre, and J. Rougemont. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in functional genomics*, 9(5-6): 466, 2010.

M. Levine. Paused RNA Polymerase II as a Developmental Checkpoint. *Cell*, 145(4): 502–511, 2011.

B. Li, L.A. Howe, S. Anderson, J.R. Yates, and J.L. Workman. The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. *Journal of Biological Chemistry*, 278(11):8897, 2003.

B. Li, M. Carey, and J.L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007.

E. Li, T.H. Bestor, and R. Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 2009.

T. Li, J.F. Hu, X. Qiu, J. Ling, H. Chen, S. Wang, A. Hou, T.H. Vu, and A.R. Hoffman. CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. *Molecular and cellular biology*, 28(20): 6473, 2008.

S.M. Liao, J. Zhang, D.A. Jeffery, A.J. Koleske, C.M. Thompson, D.M. Chao, M. Viljoen, H.J.J. van Vuuren, and R.A. Young. A kinase-cyclin pair in the RNA polymerase II holoenzyme. *Nature*, 374(6518):193–196, 1995.

Y. Liu, C. Kung, J. Fishburn, A.Z. Ansari, K.M. Shokat, and S. Hahn. Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex. *Molecular and cellular biology*, 24(4):1721, 2004.

Y.H. Loh, Q. Wu, J.L. Chew, V.B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics*, 38(4):431–440, 2006.

X. Lu, M.D. Simon, J.V. Chodaparambil, J.C. Hansen, K.M. Shokat, and K. Luger. The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. *Nature structural & molecular biology*, 15(10):1122, 2008.

R.F. Luco, Q. Pan, K. Tominaga, B.J. Blencowe, O.M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996, 2010.

A.H. Lund and M. van Lohuizen. Polycomb complexes and silencing mechanisms. *Current opinion in cell biology*, 16(3):239–246, 2004.

M. Madan Babu, S.C. Janga, I. de Santiago, and A. Pombo. Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current opinion in genetics & development*, 18(6):571–582, 2008.

Martin Maechler, Peter Rousseeuw, Anja Struyf, and Mia Hubert. Cluster Analysis Basics and Extensions. 2005.

J.R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nature genetics*, 38(10): 1151–1158, 2006.

D. Mancini-DiNardo, S.J.S. Steele, R.S. Ingram, and S.M. Tilghman. A differentially methylated region within the gene kcnq1 functions as an imprinted promoter and silencer. *Human molecular genetics*, 12(3):283, 2003.

R. Margueron and D. Reinberg. The Polycomb complex PRC2 and its mark in life. *Nature*, 469(7330):343–349, 2011.

N.F. Marshall, J. Peng, Z. Xie, and D.H. Price. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *Journal of Biological Chemistry*, 271(43):27176, 1996.

G.R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634, 1981.

G.A. Maston, S.K. Evans, and M.R. Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

A. Mattout and E. Meshorer. Chromatin plasticity and genome organization in pluripotent embryonic stem cells. *Current Opinion in Cell Biology*, 22(3):334–341, 2010.

T. Max, M. Sogaard, and J.Q. Svejstrup. Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *Journal of Biological Chemistry*, 282(19):14113, 2007.

A. Mayer, M. Lidschreiber, M. Siebert, K. Leike, J. Soding, and P. Cramer. Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology*, 17(10):1272–1278, 2010.

K.J. Meaburn and T. Misteli. Chromosome territories. *Nature*, 445(7126):379–381, 2007.

A. Meinhart, T. Kamenski, S. Hoeppner, S. Baumli, and P. Cramer. A structural perspective of CTD function. *Genes & development*, 19(12):1401, 2005.

M. Meininghaus, R.D. Chapman, M. Horndasch, and D. Eick. Conditional expression of RNA polymerase II in mammalian cells. *Journal of Biological Chemistry*, 275(32): 24375, 2000.

A. Meissner, T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein, C. Nusbaum, D.B. Jaffe, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

E.M. Mendenhall, R.P. Koche, T. Truong, V.W. Zhou, B. Issac, A.S. Chi, M. Ku, and B.E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genetics*, 6:12, 2010.

E. Meshorer and T. Misteli. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature Reviews Molecular Cell Biology*, 7(7):540–546, 2006.

T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007. ISSN 0028-0836.

T.A. Milne, Y. Dou, M.E. Martin, H.W. Brock, R.G. Roeder, and J.L. Hess. MLL associates specifically with a subset of transcriptionally active target genes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14765, 2005.

I.M. Min, J.J. Waterfall, L.J. Core, R.J. Munroe, J. Schimenti, and J.T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes & Development*, 25(7):742, 2011.

A. Missra and D.S. Gilmour. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proceedings of the National Academy of Sciences*, 107(25):11301, 2010.

J.A. Mitchell and P. Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & development*, 22(1):20, 2008.

F. Mohn and D. Schubeler. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends in Genetics*, 25(3):129–136, 2009.

F. Mohn, M. Weber, M. Rebhan, T.C. Roloff, J. Richter, M.B. Stadler, M. Bibel, and D. Schubeler. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell*, 30(6):755–766, 2008.

A.L. Mosley, S.G. Pattenden, M. Carey, S. Venkatesh, J.M. Gilmore, L. Florens, J.L. Workman, and M.P. Washburn. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Molecular cell*, 34(2):168–178, 2009.

K. Mousavi, H. Zare, A.H. Wang, and V. Sartorelli. Polycomb Protein Ezh1 Promotes RNA Polymerase II Elongation. *Molecular Cell*, 2011.

N.M. Mulholland, I.F.G. King, and R.E. Kingston. Regulation of Polycomb group complexes by the sequence-specific DNA binding proteins Zeste and GAGA. *Genes & development*, 17(22):2741, 2003.

G.W. Muse, D.A. Gilchrist, S. Nechaev, R. Shah, J.S. Parker, S.F. Grissom, J. Zeitlinger, and K. Adelman. RNA polymerase is poised for activation across the genome. *Nature genetics*, 39(12):1507–1511, 2007.

S. Nechaev, D.C. Fargo, G. Dos Santos, L. Liu, Y. Gao, and K. Adelman. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, 327(5963):335, 2010.

J.A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308, 1965.

J. Nevado, L. Gaudreau, M. Adam, and M. Ptashne. Transcriptional activation by artificial recruitment in mammalian cells. *Proceedings of the National Academy of Sciences*, 96(6):2674, 1999.

H.H. Ng, F. Robert, R.A. Young, and K. Struhl. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular cell*, 11(3):709–719, 2003.

H. Niwa, J. Miyazaki, and A.G. Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature genetics*, 24(4): 372–376, 2000.

M. Nonet, D. Sweetser, and R.A. Young. Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell*, 50(6):909–915, 1987.

J.E. Ohm, K.M. McGarvey, X. Yu, L. Cheng, K.E. Schuebel, L. Cope, H.P. Mohammad, W. Chen, V.C. Daniel, W. Yu, et al. A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics*, 39(2):237–242, 2007.

S.K.T. Ooi, C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S.P. Lin, C.D. Allis, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, 448(7154):714–717, 2007.

C.S. Osborne, L. Chakalova, K.E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J.A. Mitchell, S. Lopes, W. Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065–1071, 2004.

A. Oshlack, M.D. Robinson, and M.D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(12):220, 2010.

Z. Ouyang, Q. Zhou, and W.H. Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521, 2009.

G. Pan, S. Tian, J. Nie, C. Yang, V. Ruotti, H. Wei, G.A. Jonsdottir, R. Stewart, and J.A. Thomson. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell stem cell*, 1(3):299–312, 2007.

R.R. Pandey, T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-DiNardo, and C. Kanduri. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell*, 32(2):232–246, 2008.

P.J. Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

Y. Pei, B. Schwer, and S. Shuman. Interactions between fission yeast Cdk9, its cyclin partner Pch1, and mRNA capping enzyme Pct1 suggest an elongation checkpoint for mRNA quality control. *Journal of Biological Chemistry*, 278(9):7180, 2003.

D. Peric-Hupkes, W. Meuleman, L. Pagie, S.W.M. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R.M. Kerkhoven, M. van Lohuizen, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4):603–613, 2010.

C.L. Peterson and J.L. Workman. Promoter targeting and chromatin remodeling by the swi/snf complex. *Current opinion in genetics & development*, 10(2):187–192, 2000.

K. Pfannkuche, H. Summer, O. Li, J. Hescheler, and P. Droge. The High Mobility Group Protein HMGA2: A Co-Regulator of Chromatin Structure and Pluripotency in Stem Cells? *Stem Cell Reviews and Reports*, 5(3):224–230, 2009.

N.J. Proudfoot, A. Furger, and M.J. Dye. Integrating mRNA processing with transcription. *Cell*, 108(4):501–512, 2002.

A.R. Quinlan and I.M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841, 2010.

R Development Core Team. R: A Language and Environment for Statistical Computing. 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

P.B. Rahl, C.Y. Lin, A.C. Seila, R.A. Flynn, S. McCuine, C.B. Burge, P.A. Sharp, and R.A. Young. c-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445, 2010. ISSN 0092-8674.

V.R. Ramirez-Carrozzi, D. Braas, D.M. Bhatt, C.S. Cheng, C. Hong, K.R. Doty, J.C. Black, A. Hoffmann, M. Carey, and S.T. Smale. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell*, 138(1):114–128, 2009.

A. Rechtsteiner, S. Ercan, T. Takasaki, T.M. Phippen, T.A. Egelhofer, W. Wang, H. Kimura, J.D. Lieb, and S. Strome. The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny. *PLoS Genetics*, 6(9):e1001091, 2010.

W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425, 2007.

P. Richard and J.L. Manley. Transcription termination by nuclear RNA polymerases. *Genes & development*, 23(11):1247, 2009.

J.E. Richardson. Fjoin: simple and efficient computation of feature overlaps. *Journal of Computational Biology*, 13(8):1457–1464, 2006.

H. Richly, M. Lange, E. Simboeck, and L. Di Croce. Setting and resetting of epigenetic marks in malignant transformation and development. *BioEssays*, 32(8):669–679, 2010.

P. Rickert, J.L. Corden, and E. Lees. Cyclin C/CDK8 and cyclin H/CDK7/p36 are biochemically distinct CTD kinases. *Oncogene*, 18(4):1093, 1999.

J.L. Rinn, M. Kertesz, J.K. Wang, S.L. Squazzo, X. Xu, S.A. Brugmann, L.H. Goodnough, J.A. Helms, P.J. Farnham, E. Segal, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, 2007.

C.R. Rodriguez, E.J. Cho, M.C. Keogh, C.L. Moore, A.L. Greenleaf, and S. Buratowski. Kin28, the TFIIH-associated carboxy-terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. *Molecular and Cellular Biology*, 20(1):104, 2000.

T.Y. Roh, S. Cuddapah, K. Cui, and K. Zhao. The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences*, 103 (43):15782, 2006.

A.E. Rougvie and J.T. Lis. The RNA polymerase II molecule at the 5'end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6): 795–804, 1988.

J.M. Rozenberg, A. Shlyakhtenko, K. Glass, V. Rishi, M.V. Myakishev, P.C. FitzGerald, and C. Vinson. All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC genomics*, 9(1): 67, 2008.

AI Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374, 2003.

A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D.A. Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6):424–436, 2007.

S. Schoeftner, A.K. Sengupta, S. Kubicek, K. Mechtler, L. Spahn, H. Koseki, T. Jenuwein, and A. Wutz. Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *The EMBO Journal*, 25(13):3110–3122, 2006.

D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5): 887–898, 2008.

B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. Van Lohuizen, A. Tanay, and G. Cavalli. Functional anatomy of polycomb and

trithorax chromatin landscapes in Drosophila embryos. *PLoS biology*, 7(1):e1000013, 2009.

Y.B. Schwartz and V. Pirrotta. Polycomb complexes and epigenetic states. *Current opinion in cell biology*, 20(3):266–273, 2008.

Y.B. Schwartz, T.G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta. Alternative epigenetic chromatin states of polycomb target genes. *PLoS genetics*, 6(1):e1000805, 2010.

A.C. Seila, J.M. Calabrese, S.S. Levine, G.W. Yeo, P.B. Rahl, R.A. Flynn, R.A. Young, and P.A. Sharp. Divergent transcription from active promoters. *Science*, 322(5909): 1849, 2008.

K.H. Sene, C.J. Porter, G. Palidwor, C. Perez-Iratxeta, E.M. Muro, P.A. Campbell, M.A. Rudnicki, and M.A. Andrade-Navarro. Gene function in early mouse embryonic stem cell differentiation. *BMC genomics*, 8(1):85, 2007.

A.A. Sharov and M.S.H. Ko. Human ES cell profiling broadens the reach of bivalent domains. *Cell Stem Cell*, 1(3):237–238, 2007.

X. Shen, W. Kim, Y. Fujiwara, M.D. Simon, Y. Liu, M.R. Mysliwiec, G.C. Yuan, Y. Lee, and S.H. Orkin. Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells. *Cell*, 139(7):1303–1314, 2009.

J. Silva and A. Smith. Capturing pluripotency. *Cell*, 132(4):532–536, 2008.

M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature*, 200:6.

R.J. Sims, L.A. Rojas, D. Beck, R. Bonasio, R. Schuller, W.J. Drury, D. Eick, and D. Reinberg. The C-Terminal Domain of RNA Polymerase II Is Modified by Site-Specific Methylation. *Science*, 332(6025):99, 2011.

F. Sleutels, R. Zwart, and D.P. Barlow. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415(6873):810–813, 2002.

S.T. Smale and J.T. Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449, 2003.

M.V. Sobennikova, E.K. Shematorova, and G.V. Shpakovski. C-terminal domain of subunit Rpb1 of nuclear RNA polymerase II and its role in the transcription cycle. *Molecular Biology*, 41(3):387–401, 2007.

N. Spies, C.B. Nielsen, R.A. Padgett, and C.B. Burge. Biased chromatin signatures around polyadenylation sites and exons. *Molecular cell*, 36(2):245–254, 2009.

M. Spivakov and A.G. Fisher. Epigenetic signatures of stem-cell identity. *Nature Reviews Genetics*, 8(4):263–271, 2007.

D. Sproul, N. Gilbert, and W.A. Bickmore. The role of chromatin structure in regulating the expression of clustered genes. *Nature Reviews Genetics*, 6(10):775–781, 2005.

S.L. Squazzo, H. OGeen, V.M. Komashko, S.R. Krig, V.X. Jin, S. Jang, R. Margueron, D. Reinberg, R. Green, and P.J. Farnham. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome research*, 16(7):890, 2006.

A.L. Statham, D. Strbenac, M.W. Coolen, C. Stirzaker, S.J. Clark, and M.D. Robinson. Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics*, 26(13):1662, 2010.

E.J. Steinmetz, C.L. Warren, J.N. Kuehner, B. Panbehi, A.Z. Ansari, and D.A. Brow. Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Molecular cell*, 24(5):735–746, 2006. ISSN 1097-2765.

J.K. Stock, S. Giadrossi, M. Casanova, E. Brookes, M. Vidal, H. Koseki, N. Brockdorff, A.G. Fisher, and A. Pombo. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nature cell biology*, 9(12): 1428–1435, 2007.

A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545, 2005.

H. Sutherland and W.A. Bickmore. Transcription factories: gene expression in unions? *Nature Reviews Genetics*, 10(7):457–466, 2009.

H. Szutorisz, N. Dillon, and L. Tora. The role of enhancers as centres for general transcription factor recruitment. *Trends in biochemical sciences*, 30(11):593–599, 2005.

P. Taghavi and M. van Lohuizen. Two paths to silence merge. *Nature*, 439:16, 2006.

M. Tahiliani, K.P. Koh, Y. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer, D.R. Liu, L. Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324 (5929):930, 2009.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

H. Takeshima, S. Yamashita, T. Shimazu, T. Niwa, and T. Ushijima. The presence of RNA polymerase II, active or stalled, predicts epigenetic fate of promoter CpG islands. *Genome research*, 19(11):1974, 2009.

A. Tanay, A.H. O'Donnell, M. Damelin, and T.H. Bestor. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proceedings of the National Academy of Sciences*, 104(13):5521, 2007.

N. Thakur, V.K. Tiwari, H. Thomassin, R.R. Pandey, M. Kanduri, A. Gondor, T. Grange, R. Ohlsson, and C. Kanduri. An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region. *Molecular and cellular biology*, 24 (18):7855, 2004.

J.P. Thomson, P.J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A.R.W. Kerr, A. Deaton, R. Andrews, K.D. James, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464(7291):1082–1086, 2010.

J.R. Tietjen, D.W. Zhang, J.B. Rodriguez-Molina, B.E. White, M.S. Akhtar, M. Heidemann, X. Li, R.D. Chapman, K. Shokat, S. Keles, et al. Chemical-genomic dissection of the CTD code. *Nature Structural & Molecular Biology*, 17(99):1154–1161, 2010.

H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcarcel, and R. Guigo. Nucleosome positioning as a determinant of exon recognition. *Nature structural & molecular biology*, 16(9):996–1001, 2009.

Y. Toyooka, D. Shimosato, K. Murakami, K. Takahashi, and H. Niwa. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development*, 135(5):909, 2008.

C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105, 2009.

C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

V. Tripathi, J.D. Ellis, Z. Shen, D.Y. Song, Q. Pan, A.T. Watt, S.M. Freier, C.F. Bennett, A. Sharma, P.A. Bubulya, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell*, 2010.

J. Trott, K. Hayashi, A. Surani, M.M. Babu, and A. Martinez-Arias. Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol. BioSyst.*, 2012.

B.M. Turner. Open chromatin and hypertranscription in embryonic stem cells. *Cell Stem Cell*, 2(5):408–410, 2008.

E. Valen, G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, M. Murata, H. Nishiyori, D. Lazarevic, D. Motti, et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome research*, 19(2):255, 2009.

P. Van Der Stoop, E.A. Boutsma, D. Hulsman, S. Noback, M. Heimerikx, R.M. Kerkhoven, J.W. Voncken, L.F.A. Wessels, and M. Van Lohuizen. Ubiquitin E3 ligase Ring1b/Rnf2 of polycomb repressive complex 1 contributes to stable maintenance of mouse embryonic stem cells. *PLoS One*, 3(5):e2235, 2008.

D. Van Hoof, R. Passier, D. Ward-Van Oostwaard, M.W.H. Pinkse, A.J.R. Heck, C.L. Mummery, and J. Krijgsveld. A quest for human and mouse embryonic stem cell-specific proteins. *Molecular & Cellular Proteomics*, 5(7):1261, 2006.

S. Varambally, S.M. Dhanasekaran, M. Zhou, T.R. Barrette, C. Kumar-Sinha, M.G. Sanda, D. Ghosh, K.J. Pienta, R. Sewalt, A.P. Otte, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, 419(6907):624–629, 2002.

W.N. Venables and D.M. Smith. the R Development Core Team. *An Introduction to R, Version*, 1(0), 2003.

J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304, 2001.

M. Vermeulen, K.W. Mulder, S. Denissov, WWM Pijnappel, F. van Schaik, R.A. Varier, M. Baltissen, H.G. Stunnenberg, M. Mann, H.T. Timmers, et al. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131(1):58–69, 2007.

R. Versteeg, B.D.C. Van Schaik, M.F. Van Batenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, and A.H.C. van Kampen. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome research*, 13(9):1998, 2003.

E. Vire, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J.M. Vanderwinden, et al. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439(7078):871–874, 2006.

J.W. Voncken, B.A.J. Roelen, M. Roefs, S. De Vries, E. Verhoeven, S. Marino, J. Deschamps, and M. Van Lohuizen. Rnf2 (Ring1b) deficiency causes gastrulation arrest and cell cycle inhibition. *Proceedings of the National Academy of Sciences*, 100(5):2468, 2003.

HB Wang, L. Wang, H. Erdjument-Bromage, M. Vidal, P. Tempst, R.S. Jones, and Y. Zhang. Role of H2A ubiquitination in Polycomb silencing. *Nature*, 431:873–878, 2004.

J. Wang, S. Rao, J. Chu, X. Shen, D.N. Levasseur, T.W. Theunissen, and S.H. Orkin. A protein interaction network for pluripotency of embryonic stem cells. *Tandem*, 1 (10):30, 2006.

L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1): 136, 2010.

M. Weber, I. Hellmann, M.B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature genetics*, 39(4):457–466, 2007.

G. Wei, L. Wei, J. Zhu, C. Zang, J. Hu-Li, Z. Yao, K. Cui, Y. Kanno, T.Y. Roh, W.T. Watford, et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity*, 30(1):155–167, 2009.

M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D.J. Weisenberger, M. Campan, J. Young, I. Jacobs, et al. Epigenetic stem cell signature in cancer. *Nature genetics*, 39(2):157–158, 2006.

R.L. Williams, D.J. Hilton, S. Pease, T.A. Willson, C.L. Stewart, D.P. Gearing, E.F. Wagner, D. Metcalf, N.A. Nicola, and N.M. Gough. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature*, 336 (6200):684–687, 1988.

R.R.E. Williams, V. Azuara, P. Perry, S. Sauer, M. Dvorkina, H. Jørgensen, J. Roix, P. McQueen, T. Misteli, M. Merkenschlager, et al. Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *Journal of Cell Science*, 119 (1):132, 2006.

A. Wood and A. Shilatifard. Bur1, Bur2 and the Ctk complex in yeast: the split personality of mammalian P-TEFb. *Cell Cycle*, 5(10):1066–1068, 2006.

C.H. Wu, Y. Yamaguchi, L.R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes & development*, 17(11):1402, 2003.

H. Wu, A.C. D Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y.E. Sun, and Y. Zhang. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473(7347):389–393, 2011.

Z. Wu, R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

J. Wysocka, T. Swigut, H. Xiao, T.A. Milne, S.Y. Kwon, J. Landry, M. Kauer, A.J. Tackett, B.T. Chait, P. Badenhorst, et al. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86, 2006.

T. Xiao, H. Hall, K.O. Kizer, Y. Shibata, M.C. Hall, C.H. Borchers, and B.D. Strahl. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. *Genes & Development*, 17(5):654, 2003. ISSN 0890-9369.

S.Q. Xie, S. Martin, P.V. Guillot, D.L. Bentley, and A. Pombo. Splicing speckles are not reservoirs of RNA polymerase II, but contain an inactive form, phosphorylated on serine2 residues of the C-terminal domain. *Molecular biology of the cell*, 17(4):1723, 2006.

H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C.L. Wei, F. Lin, and W.K. Sung. A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199, 2010.

Y.X. Xu and J.L. Manley. Pin1 modulates RNA polymerase II activity during the transcription cycle. *Genes & development*, 21(22):2950, 2007.

T. Yamada, Y. Yamaguchi, N. Inukai, S. Okamoto, T. Mura, and H. Handa. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Molecular cell*, 21(2):227–237, 2006.

Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51, 1999.

G.S. Yochum and D.E. Ayer. Role for the mortality factors MORF4, MRGX, and MRG15 in transcriptional repression via associations with Pf1, mSin3A, and Transducin-Like Enhancer of Split. *Molecular and cellular biology*, 22(22):7868, 2002.

M.D. Young, T.A. Willson, M.J. Wakefield, E. Trounson, D.J. Hilton, M.E. Blewitt, A. Oshlack, and I.J. Majewski. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Research*, 2011.

C. Zang, D.E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25(15):1952, 2009.

J. Zeitlinger, A. Stark, M. Kellis, J.W. Hong, S. Nechaev, K. Adelman, M. Levine, and R.A. Young. RNA Polymerase Stalling at Developmental Control Genes in the Drosophila Embryo. *Nature genetics*, 39(12):1512, 2007.

P. Zhang, J. Du, B. Sun, X. Dong, G. Xu, J. Zhou, Q. Huang, Q. Liu, Q. Hao, and J. Ding. Structure of human MRG15 chromo domain and its binding to Lys36-methylated histone H3. *Nucleic acids research*, 34(22):6621, 2006a.

Y. Zhang, Y. Kim, N. Genoud, J. Gao, J.W. Kelly, S.L. Pfaff, G.N. Gill, J.E. Dixon, and J.P. Noel. Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Molecular cell*, 24(5):759–770, 2006b.

J. Zhao, B.K. Sun, J.A. Erwin, J.J. Song, and J.T. Lee. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322(5902):750, 2008.

X.D. Zhao, X. Han, J.L. Chew, J. Liu, K.P. Chiu, A. Choo, Y.L. Orlov, W.K. Sung, A. Shahab, V.A. Kuznetsov, et al. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell stem cell*, 1(3):286–298, 2007.

W. Zhou, P. Zhu, J. Wang, G. Pascual, K.A. Ohgi, J. Lozach, C.K. Glass, and M.G. Rosenfeld. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Molecular cell*, 29(1):69–80, 2008.

D. Zilberman, D. Coleman-Derr, T. Ballinger, and S. Henikoff. Histone H2A. Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218): 125, 2008.