Imperial College London

Unravelling the proteome of chromatin bound RNA polymerase II using Proteome-ChIP in murine stem cells

Thesis submitted for the degree of Doctor of Philosophy by

Kedar Nath Natarajan

Department of Mathematics

MRC Clinical Sciences Centre, London

April 2013

Declaration of Originality

I hereby declare that the thesis 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, April 2013

_____

Kedar Nath Natarajan

Abstract

Regulation of gene expression is critical to govern distinct transcriptional programs for a cell type, lineage specification and developmental stage. Transcription is the first step in gene expression wherein RNA Polymerase II (RNAPII) transcribes protein-coding genes. Transcription is a highly coordinated process that involves a range of chromatin interactions including transcription machinery, chromatin remodellers and co-transcriptional RNA processing. Embryonic stem (ES) cells are pluripotent, self-renewing cells that can differentiate to give rise to all lineages making them an invaluable tool to study early development and in therapy. Genome-wide analysis in murine mES cells has identified 30% of known genes harbouring bivalent chromatin modifications along with repressive Polycomb complexes and a novel variant of RNAPII (modified as $S5p^{+}S7p^{-}S2p^{-}$) with mechanistic implications in stem cell pluripotency, differentiation potential and lineage specification.

To explore chromatin composition associated with different variants of RNAPII, I developed an unbiased method, 'Proteome-ChIP' (pChIP) wherein crosslinked chromatin is purified by immunoprecipitation followed by protein extraction and identification by Mass Spectrometry. Using an unbiased comprehensive experimental strategy and a novel systems biology approach, I qualitatively and quantitatively dissect the proteome composition and dependencies on RNAPII modifications during different stages of the transcription cycle. The work done in this thesis provides an invaluable resource of RNAPII chromatin interactions. We identify known and novel components of the co-transcriptional machinery, chromatin remodelling and RNA processing machinery. The work also uncovers novel processes associated with unusual RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) including DNA replication, Polycomb proteins and chromatin remodellers; many of these processes critical for stem cell viability and regulation.

Extending the RNAPII-pChIP analysis on low complexity samples by Native-pChIP and Gradient-pChIP highlights the versatility of robustness of our method. The work described in this sheds light on regulatory chromatin processes specific to mES cells, which informs our understanding of stem cell biology and reprogramming.

Acknowledgment

Firstly I would like to express my heartfelt thanks to my supervisor and mentor Ana Pombo, for her guidance, encouragement and kindness. She is an inspiration and I am in her gratitude for her enthusiasm, advice and support in shaping up my scientific career and throughout the PhD journey. Thank you again for taking always making the time and creating a wonderful atmosphere to work.

I would like to specially thank my co-supervisor Mauricio Barahona for his encouragement and support during the PhD and for the wonderful collaboration. I would also like to thank Borislav Vangelov for discussions and producing eye-catching network images.

Thank you to Bram Snijders for the wonderful Mass spectrometry, many fruitful discussions and support during the course of this work. Having the mass spectrometry starting at the same time as my PhD and on the same floor was truly opportune.

For all the help, laughter and friendship, special mention to past and current member of Genome Function laboratory, thanks to Andre, Emily, Mita, Claudia, Ines Santiago, Liron, Sheila, Carmelo, Kelly, Ines Castro, Joao and Rob.

I am grateful to Marie-Curie (FP7), MRC and Imperial College for providing funds and support during this thesis. I am most thankful to all the wonderful FP7 people who have made this journey exciting and for laughs during our meetings. Thank you Marion, Viola, Yuva, Giorgio, Sypros and many others for always being there and for fun times. I am thankful to FP7 coordinator John Strouboulis and other Marie-Curie PI's for their guidance and advice.

Thanks for many people at the CSC who have helped in variety of ways. Thanks to my good friend Vineet for helping me transition in London and 15minutes of tea in

evenings. Thanks to the Indian tea-coffee team including Prashant, Gopu, Sanjay, Preksha and Shichina.

Thanks also to all contributors and collaborators in this project, in particular to those whose work I present I this thesis. Thank you for the advice, contribution and stimulating discussions. I would also like to sincerely thank all the people who have helped in different ways but have not been mentioned individually.

Finally sincere and enormous thanks to my fantastic family for their consistent support, encouragement and blessings. I owe so much to them. To my nephew and niece that bring joy and smile every time. I am most thankful to my father and brothers, who have supported me through everything, always believed in me and pushed me, achieve anything I set my goals to!

Table of contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

List of tables

Abbreviations

ABL	Abelson tyrosine kinase
ab	antibody
ac	acetylation
AP	alkaline phosphatase
ATP	adenosine tri-phosphate
Bmi1	B lymphoma Mo-MLV insertion region 1
BRD4	bromodomain-containing protein 4
BSA	bovine serum albumin
CAK	CDK-activating kinase
CDK	cyclin dependent kinase
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
CKII	casein kinase II
CTD	carboxy-terminal domain
CUT	cryptic unstable transcripts
Dig	digoxigenin
DNase	deoxyribonuclease
DNA	deoxyribonucleic acid
DRB	5,6-dichloro-1-b-D-ribofuranosylbenimidazole
DSIF	DRB-sensitivity inducing factor
EDTA	ethylenediaminetetraacetic acid
Eed	embryonic ectoderm development
ERK	extracellular-regulated kinase
ES	embryonic stem
Ezh2	enhancer of zeste homolog 2
FCP1	transcription factor II*F*-associated CTD *p*hosphatase 1
FCS	fetal calf serum
FGF	fibroblast growth factor
G	guanosine
GO	gene ontology

GTF	general transcription factors
H	histone
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HMT	histone methyl transferase
Hox	homeobox
hsp	heat shock protein
ICM	inner cell mass
JMJD3	Jumonji domain-containing protein 3
ICM	inner cell mass
Ig	immunoglobulin
IP	immunoprecipitation
K	lysine
LIF	leukemia inhibitory factor
LINE	long interspersed nuclear element
m7G	7-methyl guanosine
me	methylation
mRNA	messenger RNA
miRNA	micro RNA
ncRNA	non-coding RNA
MN	micrococcal nuclease
NEAA	non-essential amino acids
NEDD4	neural precursor cell expressed, developmentally down-regulated 4
NELF	negative elongation factor
Oct4	octamer-4
O-GlcNAc	O-linked N-acetylglucosamine
OGT	O-linked N-acetylglucosamine transferase
p	phosphorylation
P	proline
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline

Pcl	Polycomb-like
PCR	polymerase chain reaction
Pho	pleiohomeotic
PIC	pre-initiation complex
PIN1	protein interacting with NIMA (never in mitosis A)-1
PMSF	phenylmethyl-sulphonyl fluoride
polyA	polyadenylation
PRC	Polycomb repressive complex
PRE	Polycomb response element
PRCi	PRC-intermediate cluster
PRCo	PRC-only cluster
PRCr	PRC-repressed cluster
Psc	posterior sex combs
P-TEFb	positive transcription elongation factor b
RA	retinoic acid
Rpb1	RNA polymerase II subunit B1
REST	RE-1 silencing transcription factor
Ring	really interesting new gene
RNA	ribonucleic acid
RNAP	RNA polymerase
RNAPII S2p	RNAPII phosphorylated on serine 2 residues of the CTD
RNAPII S5p	RNAPII phosphorylated on serine 5 residues of the CTD
RNAPII S7p	RNAPII phosphorylated on serine 7 residues of the CTD
RNAPII Y1p	RNAPII phosphorylated on tyrosine 1 residues of the CTD
RNAPII T4p	RNAPII phosphorylated on threonine 4 residues of the CTD
rRNA	ribosomal RNA
Rpap	RNA polymerase associated protein
rpm	revolutions per minute
RT	real-time
Rtr	regulator of transcription
S	serine
Scp	small CTD phosphatase

Seq	sequencing
Set	suppressor of variegation, enhancer of zeste and trithorax
Setdb1	SET domain bifurcated 1
SINE	short interspersed nuclear element
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
Sox2	SRY (sex determining region Y)-box 2
SPT	suppressor of Ty
Ssu72	suppressors of *sua*7 protein 2
SUMO	small ubiquitin-related modifier
Suz12	suppressor of zeste 12
SWI/SNF	switch/sucrose non-fermentable
UTX	ubiquitously transcribed tetratricopeptide repeat, X chromosome
T	threonine
T	thymidine
TBP	TATA-binding protein
TE	trophectoderm
TF	transcription factor
TFIIH	transcription factor II H
tRNA	transfer RNA
TS	trophoblast stem
TSS	transcription start site
ub	ubiquitination
USP	ubiquitin-specific protease
UTR	un-translated region
UV	ultraviolet
WT	wild-type
WWP2	WW domain containing E3 ubiquitin protein ligase 2
XEN	extra-embryonic endoderm
Y	tyrosine
YY1	yin-yang-1

1. Literature review and thesis overview

1.1. Gene regulation and control of gene expression

Every major developmental process is driven by changes in the pattern of gene expression. This tightly regulated dynamic process governs distinct transcriptional programs whereby genes are activated or silenced, conferring specificity to a cell type, to its distinct stage in lineage specification and to developmental stage. Temporal control of gene expression allows cells to remain responsive to external cues (including developmental or environmental) and mediate appropriate changes in transcription. Genome regulation and control of gene expression can be observed, visualised and understood at different scales of genome resolution. These include from the large scale positioning of chromatin to the local chromatin regulation effects: chromosomal associations with nuclear landmarks, long-range chromatin interactions, recruitment of transcription factors, master regulators, chromatin remodellers, sequence specific factors and the transcription and co-transcriptional RNA processing machinery. Transcription is the first step in gene expression and a major target of regulation; establishment of a permissive chromatin architecture allows RNA polymerases to transcribe the genome, producing RNA molecules that, after complex co-transcriptional processing, act as template message for translation to protein product or serve independently with its structural and functional roles. Thus, the interactions at different levels allow us to understand the full range of complexities that result in transcriptional control.

1.1.1. Genome and its packaging.

In three-dimensional nuclei of each cell of an organism, the genetic information that specifies protein instruction for functioning and development is encoded in the DNA. The DNA is highly folded, constrained and compacted at several levels and the dynamics of higher-order structures play crucial roles in transcription and biological processes inherent to DNA.

Histone and non-histone proteins hierarchically package the genomic DNA into chromatin in the eukaryotic nucleus. The first level of chromatin organisation involves packaging of DNA around octamer of histone proteins forming nucleosome monomer. Packaging of nucleosomal array in a 'beads-on-a-string' conformation and binding with linker histone (H1 or H5) further organizes the nucleosome arrays into a more condensed chromatin fibre and finally into chromosomes (Jenuwein and Allis 2001; Kouzarides 2007).

At the sequence level, DNA features like gene density, base composition, CpG levels, intron/exon density, repeats and motifs, all influence gene regulation and further coordinate with enhancers and insulators to influence the transcriptional status. In addition, epigenetic modifications that heritably influence gene expression without affecting the underlying DNA sequence layer additional levels of complexity and regulation. The local chromatin architecture composed of underlying DNA sequence with features and epigenetic modifications dynamically facilitate binding of transcription factors, chromatin remodellers and master regulators thereby regulating the transcriptional states. Other mechanisms that influence and govern transcription include histone modifications, nucleosome positioning, higher-order chromatin structure and genome architecture inside the three-dimensional nucleus (Strahl and Allis 2000; Branco and Pombo 2007; Segal and Widom 2009; Bannister and Kouzarides 2011).

1.1.2. Chromatin and epigenetic regulation

The chromatin inside cells is a diverse molecular ensemble consisting of DNA bound together with all directly or indirectly associating proteins or RNA molecules. These include histones, DNA-binding factors, RNA molecules, transcriptional machinery, co-transcriptional factors and nascent transcripts, replication and repair machineries that copy and maintain DNA molecules, and regulate their interactions. Remarkably, all the components on chromatin are thought to act in a concerted fashion responding to local architectural cues (van Steensel 2011).

Chromatin consists of nucleosomes, in which DNA (147bp) is wrapped around a protein octamer of histones composed of two subunits of H2A, H2B, H3 and H4. The intervening, linker DNA (20-50bp) can also associate with the nucleosome through interactions with linker histone H1 that sits on top of nucleosomal bead. Histones and their post-translational modifications have structural roles that control the chromatin topology and compaction, but also help recruit chromatin remodellers that utilize the energy derived from the hydrolysis of ATP to reposition nucleosomes. In addition, these modifications can influence transcription and other DNA processes occurring on chromatin such as repair, replication and recombination. Histone acetylation is regulated by histone acetyltransferases (HAT) and histone deacetylases (HDAC). Phosphorylation of histone tails is quite dynamic occurring on Serine, Threonine and Tyrosine residues and catalysed reversibly by kinases and phosphatases. Methylation of histones is better understood and occurs on Lysine and Arginine residues. Lysine methylation is reversible and regulated by activity histone methyltransferases (HKMT) and histone demethylases (Bannister and Kouzarides 2011; Follmer *et al.* 2012; Leeb and Wutz 2012; Luis *et al.* 2012). In this thesis, I will predominantly discuss active histone modifications H3K4me3, H3K36me3 enriched at actively, transcriptionally permissive chromatin and Polycomb-dependent repressive histone modifications H3K27me3 and H2Aub associated at silenced loci (Wang *et al.* 2004; Mikkelsen *et al.* 2007).

1.1.3. Chromatin proteome

The abundance and diversity of chromatin proteins is quite remarkable. For example, a human cell is thought to contain 10^{10} protein molecules; assuming 10% nuclear distribution. We are still left with 10^9 molecules that in a nucleus consist of 10^3-10^5 molecules of transcription factors, histones, high-mobility group proteins and components of transcription and co-transcriptional RNA processing (van Steensel 2011).

It has also been reported that human nuclei contain roughly 8000 different proteins and ~1400 DNA binding factors (Vaquerizas *et al.* 2009). Considering

the abundance and diversity of proteins, their respective interactions are also amplified with suggestion that the human interactome may contain 130 000 interactions which would include protein-protein, protein-RNA and protein-DNA (Venkatesan *et al.* 2009).

The myriad of chromatin proteins and their interactions that regulate chromatin have led to identification of distinct chromatin states (van Steensel *et al.* 2001; Bernstein *et al.* 2006; Brookes *et al.* 2012), but inevitably depends on the *a priori* knowledge of chromatin marks and known factors for which probes, such as antibodies, have already been developed. Mapping of all possible components bound to chromatin at the different states is challenging and expensive; moreover it does not inform about which proteins simultaneously bind to chromatin and functionally interplay. With the aim of identifying the proteome of specific chromatin states, it was my aim to develop a novel unbiased method that identifies and dissects the chromatin-bound proteome associated with a protein of interest.

1.1.4. Transcription, RNA polymerases and co-transcriptional regulation

Transcription is the process where the genetic information encoded in the DNA is transcribed into RNA by large molecular subunit machines called RNA polymerases (RNAP; (Fuda *et al.* 2009). Eukaryotes have three distinct nuclear RNAPs with each transcribing a separate set of genes (Roeder 2005). RNAPI (RNA polymerase I) transcribes abundant ribosomal RNAs (rRNA) encoded on a multicopy single gene, 45S rRNA, and RNAPIII (RNA polymerase III) transcribes structural RNAs such as transfer RNA (tRNA), 5S rRNA and U6 spliceosomal small nuclear RNA (snRNA) (Dieci *et al.* 2007; Pagano *et al.* 2007). RNAPII (RNA polymerase II) transcribes all protein-coding genes and many structural and non-coding RNAs including snRNA, small nucleolar RNA (snoRNA), microRNA (miRNA) precursor and cryptic unstable transcripts (CUTs). The process of transcription consists of three steps: initiation, elongation and termination that are highly coordinated and involving co-association with a range of co-factors.

RNAPII transcription at individual genes is highly modulated and specific regulation is critical for the development of an organism. Transcriptional regulation is achieved by a complex combinatorial set of molecular interactions occurring on chromatin between RNAPII, co-factors, specific DNA sequences and additional machineries that inherently are regulated by local chromatin architecture. Some examples of transcriptional machinery and interactions include general transcription factors (Gtfs) that facilitate recruitment of RNAPII on DNA, sequence-specific transcription factors, regulatory co-factors, Mediator complex, Integrator complex and co-transcriptional processing machinery. Co-transcription and post-transcriptional regulatory mechanisms include maturation and processing of RNA to produce stable mRNA, its export from the nucleus by specific transport factors and finally localisation of sites of translation. Gene expression is additionally regulated by components at different steps including mRNA levels and feedback, RNA interference (small interfering RNA, microRNA, etc.), long non-coding RNA and a variety of other mechanisms (Chen and Carmichael 2010; Lee 2012).

1.1.5. Gene regulation in mES cells

Embryonic stem (ES) cells are pluripotent, self-renewing cells that are derived from the inner cell mass (ICM) of the developing blastocyst in the early embryo. Pluripotency is the capacity of a single cell to generate all cell lineages of the developing adult organism and ES cells inherit this property from the ICM that *in vivo* goes on to form all cells of the proper embryo. Self-renewal is the ability of a cell to proliferate in the same state and indefinitely in culture (Young 2011; Fong *et al.* 2012; Li *et al.* 2012). These properties of ES cells makes them an invaluable tool to study early development in organisms and understand the control mechanism associated with defects in development and disease that open the horizon to devise cell therapeutic possibilities.

The ES cell transcriptome is regulated by a gene expression network that allows them to self-renew while retaining the potential to differentiate into essentially all lineages upon appropriate cues. The complex chromatin network of ES cells is regulated by master transcription factors (including Oct4, Sox2 and Nanog), their regulatory circuit, chromatin remodellers, RNAPII transcription and DNA-binding factors that cascade a plethora of downstream pathways (Orkin *et al.* 2008; Young 2011). Owing to its dynamic and complex gene regulatory network, the ES cell chromatin maintains open, more permissive state allowing activation of specific combination of genes essential for housekeeping functions and importantly the ES cell state while repressing lineage specific genes. Open chromatin architecture of ES cells contributes to its plasticity (and hyperactive transcription) and undergoes a rapid shift upon cell commitment to suit the needs to specific lineage specification (Efroni *et al.* 2008; Jaenisch and Young 2008).

The mES cell core circuitry consists of master transcription factors (Oct4, Sox2 and Nanog), their transcriptional network and protein interaction network that control the pluripotency in mES cells (Wang *et al.* 2006; Orkin *et al.* 2008; Wang and Orkin 2008). These master regulators co-occupy sites of other essential transcription factors and signalling pathways (including LIF, Wnt, Bmp4, Stat3, Smad1) thereby allowing direct control of genes to regulate their downstream transcription and maintain chromatin states (Chen *et al.* 2008a; Chen *et al.* 2008b). Specific analysis of mES cell proteins and their association with the core pluripotency network has further unravelled novel roles in pluripotency and mES cell survival.

In this project and thesis, I will focus on regulation of RNAPII in mES cells in particular understanding and unravelling the chromatin state at developmental regulator genes (under the control of a master regulator) and with aim of identifying novel protein associations in mES cell chromatin.

1.2. The RNA Polymerase II complex (RNAPII)

RNAPII protein transcribes protein-coding genes and many non-coding RNAs on chromatin and coordinates a cascade of interactions with a range of nuclear processes including transcription factors, chromatin remodellers, co-transcriptional machinery and other factors to mediate the proper control of gene expression allowing cells to grow, divide and respond appropriately to environmental and developmental cues.

1.2.1. Core subunits

RNAPII is a multi-subunit protein complex composed of 12 subunits (Rpb1-12 or Polr2a-l). Five of these subunits are shared with other RNA polymerases (Polr2e, Polr2f, Polr2h, Polr2k, Polr2l). Regulation of RNAPII and control of gene expression is mediated by a complex set of post-translational modifications that occur on the C-terminal domain (CTD) of the Rpb1 (Polr2a), the largest subunit of RNAPII. Modifications of the RNAPII integrate transcriptional process with a multitude of co-regulatory processes catalysed by a diverse range of modifying proteins.

1.2.2. Rpb1 and CTD

The largest subunit of RNAPII, Rpb1, has a unique highly repetitive CTD consisting of multiple tandem hepta-peptides that plays a central role in complex regulation of gene expression. The evolutionary conserved consensus motif of the CTD consists of $Y_1 – S_2 – P_3 – T_4 – S_5 – P_6 – S_7$ (Tyrosine – Serine – Proline – Threonine – Serine – Proline – Serine; Fig. 1.1) (Corden *et al.* 1985). CTD length differs depending on the complexity of organism. In humans and murine cells the CTD sequence is repeated 52 times, 44 times in *D. melanogaster*, 26 times in yeast, wherein all repeats almost obey consensus (Egloff and Murphy 2008b; Egloff and Murphy 2008a; Heidemann *et al.* 2012). Deletion of the CTD is lethal in yeast, *D. melanogaster* and in mouse, however mutations have revealed a dispensable role for CTD in transcription (Meininghaus *et al.* 2000).

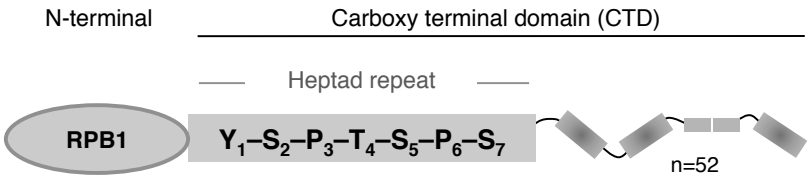


Figure 1.1 Simplified representation of Rpb1 subunit and its CTD. The largest subunit of RNAPII, Rpb1 is composed of CTD consisting or multi-heptad repeat of above consensus sequence. Number of tandem hepta-peptide repeats varies with the complexity of the organism with mouse and human protein having 52 repeats. Distal part of CTD contains amino acids that can diverge from consensus, including lysine, arginine and threonine residues.

The non-consensus repeats of the CTD (31 repeats) are mostly distally placed at C-terminal part of CTD and with changes at positions 3,4,5 and 7. Serine at position 7 is most substituted and interestingly lysine residues replace the serine residues at distal repeats (position 7). Theoretically, the presence of lysine at the CTD allows digestion with trypsin and MS detection of peptides. However, remarkably, the structure of CTD and complex modification of the heptad repeats makes detection of peptides increasingly difficult. From purified RNAPII protein (samples kindly provided by Dr. Andre Moeller; our laboratory) and using MS analysis after trypsin digestion, we observed all predicted CTD residues as listed in Table 1.1. The last CTD peptide contains an unusual extension that includes unique constitutive binding site for phosphorylated casein kinase II site (CKII) and binding sites for tyrosine kinases c-abl1, a-abl2 (Baskaran *et al.* 1996; Baskaran *et al.* 1997; Chapman *et al.* 2004). Remarkably, removal of this domain impairs thr stability of Rpb1, cellular viability and nascent RNA processing (Fong *et al.* 2003; Chapman *et al.* 2004).

Table 1.1 List of Rpb1-CTD peptides.

DNA-directed RNA polymerase II subunit RPB1	Coordinates	Length
Polr2a	1 – 1970	1970
Repeat 52; approximate	1954 – 1960	7
Region (52 X 7 AA approximate tandem repeats of Y-[ST]-P-[STQ]-[ST]-P-[SRNTEVKGN]52 X 7 AA approximate tandem repeats of Y-[ST]-P-[STQ]-[ST]-P-[SRNTEVKGN])	1593 – 1960	368

CTD starts from 1593 Sequence	Length
YGMEIPTNIPGLGAAGPTGMFFGSAPSPMGGISPAMTPWNQGATPAYGAWSPSVGSGMTPGAAGFS PSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPSYSPTSPAYEPR	122
SPGGYTPQSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPS YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTS PSYSPTSPNYSPTSPNYTPTSPSYSPTSPSYSPTSPNYTPTSPNYSPTSPSYSPTSPSYSPTSPSYSP SSPR	207
YTPQSPTYTPSSPSYSPSSPSYSPTSPK	28
YTPTSPSYSPSSPEYTPASPK	21
YSPTSPK	7
YSPTSPTYSPTSPVYTPTSPK	21
YSPTSPTYSPTSPK	14
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN	34

The Rpb1-CTD is thought to act as a binding scaffold for a variety of nuclear proteins including histone modifications, RNA processing and components of co-transcriptional machinery. Since the bound factors physically associate with RNAPII, their processes inherently become linked to transcription (Phatnani and Greenleaf 2006). In addition, the pattern of modifications on the CTD couple the recruitment and binding of different protein factors on chromatin with RNAPII. These modifications include phosphorylation, isomerisation, methylation, ubiquitination and glycosylation.
The enormous combinatorial potential of post-translational modifications on the Rpb1-CTD could allow existence of several individual RNAPII's (with distinct combinations) linked to several stages of transcription and RNA processing in living cells. This has led to the hypothesis of the CTD code (Kelly *et al.* 1993; Buratowski 2003; Egloff and Murphy 2008a)

The structure of the CTD is a largely disordered, flexible and free hanging, which enables interaction with a variety of proteins. The combination of post-translational modifications occurring on the CTD is thought to alter its 3D conformation thereby allowing accessibility of DNA binding sites and CTD to recruit additional factors. The free structure (linearized; 375 amino acids) of the Rpb1-CTD is likely to extend several diameters of the RNAPII globular structure and owing to its long structure, it can provide a surface for

simultaneous binding of several co-factors, RNA and interaction with chromatin.

1.2.3. CTD modifications

1.2.4. Serine 5 phosphorylation

Phosphorylation of serine at position 5 of the Rpb1-CTD heptad (S5p) was initially identified at promoter regions (5' end) of actively transcribing protein-coding genes that suggested a role in transcription initiation (Komarnitsky *et al.* 2000). This modification (S5p) is thought to allow RNAPII promoter escape, dissociation from the pre-initiation complex and recruitment of appropriate factors for proper transcription initiation.

The general transcription factor TFIIH subunit Cdk7 (kin28 in yeast), mediates phosphorylation of S5p. The CDK-activating kinase (CAK) is formed by association of Cdk7 with 'cyclin-H' and 'Mat1', which is incorporated into the TFIIH complex and further phosphorylates RNAPII-associated with DNA. Mediator subunit Cdk8 (Srb10 in yeast) is another enzyme that can phosphorylate S5p and has roles in gene activation (Galbraith *et al.* 2010) and also repressive roles (Hengartner *et al.* 1998).

The CTD containing S5p is thought to be landing pad that recruits and activates the capping enzyme to add cap structure to 5' end of the newly synthesised RNA. S5p is found at promoters of mRNA and snRNA genes. Mutation of S5 residue to Alanine (non-phosphoacceptor) causes a drastic reduction in steady state levels of RNA (Egloff *et al.* 2007), and is thought to be an essential step for protection of 5' end of RNA from exonucleases (Egloff and Murphy 2008b). Along with recruitment of capping enzyme, S5p is also required for interaction with Nrd1 which mediates 3'-end formation and premature termination at non-polyadenylated transcripts in yeast (Heidemann *et al.* 2012). S5p also links histone modifications and chromatin remodellers. S5p interacts with Set1 (histone methyltransferase) that catalysis H3K4me3; a mark of open chromatin and transcription initiation. In addition, S5p is also

involved in recruitment of Rpd3C, histone H3 and H4 deacetylase (Govind *et al.* 2010).

Various enzymes are involved in removing the S5p mark from RNAPII-CTD. Ssu72 (suppressor of sua7 protein) is known to dephosphorylate S5p and has additional roles in mRNA elongation and termination. Ssu72 is the primary S5p phosphatase in yeast and its activity is regulated through proline isomerisation (at positions 3 and 5 of Rpb1-CTD) (Krishnamurthy *et al.* 2009). Scp1-3 (small CTD phosphatases) are known to cooperate with REST chromatin remodelling complex and dephosphorylate S5p at REST-target genes (Yeo *et al.* 2005). Scps act as transcription regulators, silencing neuronal expression in non-neuronal genes and have also been hypothesised to play a role in general transcription (Yeo *et al.* 2005; Heidemann *et al.* 2012). Rtr1 (regulator of transcription) is an atypical S5p phosphatase identified in yeast. Rtr1 deletion in yeast is not lethal and the lack of an active phosphatase site in Rtr1 has suggested its non-catalytical role in dephosphorylation. Rtr1 deletions also lead to high S5p levels during elongation (Mosley *et al.* 2009). Rpap2 (RNA-associated protein) is the human homolog of Rtr1 and is also associated with S5p dephosphorylation. Rpap2 is detected at 5' ends of actively transcribed genes and is essential for appropriate 3' end formation of snRNA transcripts (Egloff *et al.* 2007).

1.2.5. Serine 7 phosphorylation

Phosphorylation of serine at position 7 (S7p) occurs both in yeast and in mammalian cells (Chapman *et al.* 2007; Egloff *et al.* 2007; Kim *et al.* 2010; Hajheidari *et al.* 2012). S7p is identified at promoters and coding regions of protein-coding genes and snRNA genes. Substitution of S7 to alanine (non-phospho acceptor) is not lethal is yeast but impairs viability of human cells. Mutations to the S7p residue donot affect the expression of protein-coding genes, however the drastic effect is observed on transcription and processing of snRNA genes (Egloff *et al.* 2007). snRNAs contain a conserved 3' box for

3’ end RNA processing recognised by the integrator complex and this, in association with Rtr1/Rpap2, facilitates interactions with PTEF-b and appropriate transcription elongation and snRNA processing (Egloff *et al.* 2007; Heidemann *et al.* 2013). This has led to the suggestion of CTD modifications having gene-specific functions.

In protein coding genes, S7p is placed early in transcription and the modification can be detected on RNAPII until polyadenylation site. S7p is phosphorylated by Cdk7 (kin28 in yeast), the S5p kinase, and chemical inhibition of cdk7 results in a drastic decrease of both S5p and S7p at 5’ regions of genes. Interestingly, S7p is though to influence the extent of S5p and both the marks are added synchronously during thr early stages of transcription. The role of S7p is thought as a transition from S5p (initiation) to S2p (elongation). Additional S7p kinases have been suggested, including Bur1 (yeast internal S7p kinase) (Boeing *et al.* 2010; Zhang *et al.* 2012b). In contrast to alanine mutants, S7 to glutamate (phospho-mimic) substitutions are lethal in human and in yeast. Owing to the distribution of S5p and S7p modifications and their catalysis by cdk7, the only known phosphatase acting on S7p is Ssu72 that removes S7p immediately after the polyadenylation site and reconstitutes RNAPII to hypo-phosphorylated state (Bataille *et al.* 2012; Zhang *et al.* 2012a).

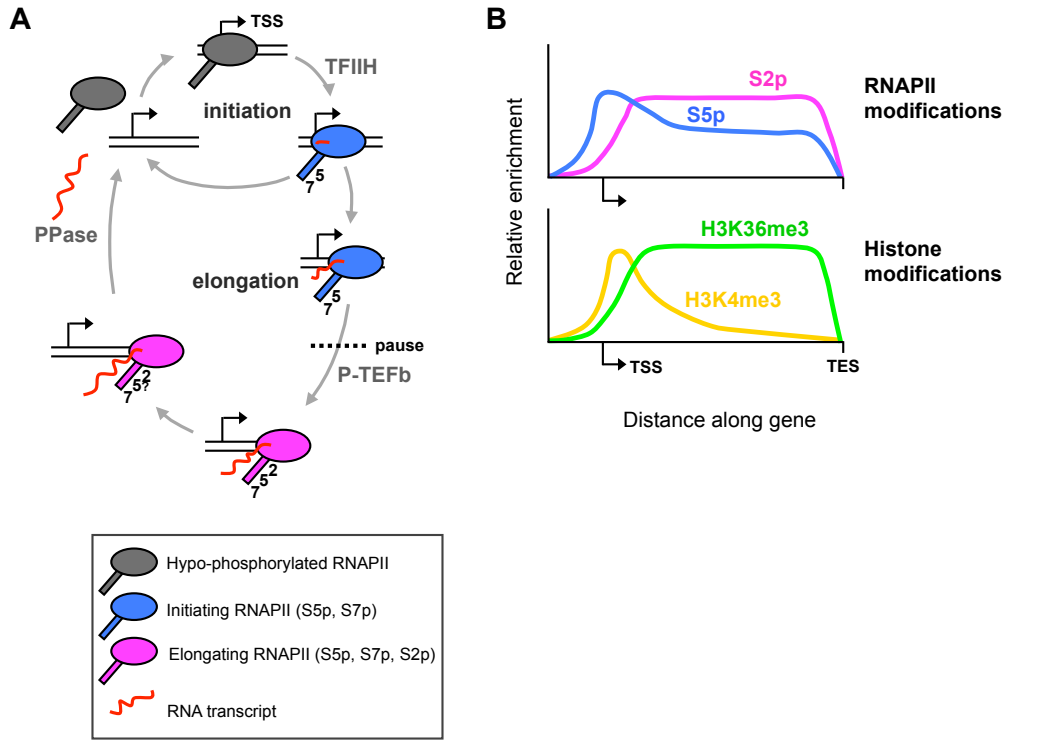


Figure 1.1 Diagrams representing phosphorylation events on Rpb1-CTD during the transcription cycle. (A) RNAPII is recruited at the gene promoters in hypo-phosphorylated form, when it undergoes modification at S5p that initiates the RNAPII. Modification at S7p and subsequently at S2p accompanies the transition of RNAPII from initiation to elongation further producing stable mRNA for export, while the RNAPII is de-phosphorylated and recycled. (B) A simplified pattern of RNAPII occupancy and histone modifications across a typical active gene.

1.2.6. Serine 2 phosphorylation

Phosphorylation of S2 (S2p) of Rpb1-CTD is identified at coding regions of actively transcribing genes and is associated with productive elongation (Komarnitsky *et al.* 2000; Morris *et al.* 2005) and recruitment of chromatin remodellers (for H3K36me3; hallmark of elongation), RNA processing factors, co-transcriptional machinery and splicing factors (Kim *et al.* 2002; Proudfoot *et al.* 2002; Li *et al.* 2005).

Cdk9 (Bur1 and Ctk1 in yeast) phosphorylates Serine 2 of the RPb1-CTD (S2p). Cdk9 is also the catalytic subunit of the PTEF-b complex (positive transcription elongation factor). The Spt5 subunit of DSIF (DRB sensitivity-inducing factor) is a non-CTD substrate of Bur1/Cdk9 containing a C-terminal repeat region (CTR). DSIF complexes with NELF (negative regulation of

transcription) and phosphorylation of this complex by Cdk9 (PTEF-b) and S2p allow progression in transcriptional elongation (Bartkowiak *et al.* 2011; Nechaev and Adelman 2011). PTEF-b is also recruited by sequence-specific TFs, such as TNFα and c-Myc (Rahl *et al.* 2010). Additionally, BRD4 interacts with P-TEFb, recruiting the kinase activity to sites of acetylated histones (Jang *et al.* 2005). More recently, other S2p kinases have been identified in higher eukaryotes. Cdk12 and Cdk13 are important for human genes, additionally Brd4 (bromodomain protein) has been shown to be atypical S2p CTD kinase *in vitro* and *in vivo* (Bartkowiak *et al.* 2010; Devaiah *et al.* 2012). S2p is thought to recruit Set2 methyltransferase, that catalyses H3K36me3 that marks transcriptionally active, elongating genes and inhibits inappropriate transcription from cryptic promoters through recruitment of repressive Rpd3s histone deacetylase complex (Carrozza *et al.* 2005; Keogh *et al.* 2005; Kizer *et al.* 2005). S2p is also required for global and gene-associated levels of histone H2B ubiquitination and correct 3' processing of replication dependent histone mRNA (Pirngruber *et al.* 2009).

Removal of S2p from the Rpb1-CTD is catalysed by evolutionarily conserved FCP1 (TFIIH associating C-terminal domain phosphatase). Fcp1 is thought to travel along with RNAPII till the end of the gene and remove S2p allowing the next round of transcription (Cho *et al.* 2001; Ghosh *et al.* 2008). Cdc14 is another phosphatase that is required for S5p and S2p dephosphorylation during mitosis and inhibition of sub-telomeric elements (Clemente-Blanco *et al.* 2011). Substitution of S2 to Alanine in Rpb1-CTD leads to defects in 3' end processing of genes consistent with roles in processing and maturation of RNA and transcriptional regulation(Medlin *et al.* 2005; Egloff *et al.* 2007).

1.2.7. Other CTD modifications

The highly dynamic nature of CTD modifications and combination of potential post-translational modifications allow existence of several RNAPII molecules each with distinct combinatorial code and this hypothesis has led to the suggestion of CTD code (Buratowski 2003; Egloff and Murphy 2008a)

The other phosphorylation on the CTD include Tyrosine phosphorylation (Y1p) (Baskaran *et al.* 1993; Baskaran *et al.* 1997; Baskaran *et al.* 1999) and Threonine phosphorylation (T4p) (Hsin *et al.* 2011; Hintermair *et al.* 2012). The tyrosine kinases include c-Abl1 and c-Abl2 and these modifications are known to increase transcription elongation rates. Threonine phosphorylation is present in gene bodies and substitution to Alanine (or valine) is non-lethal in yeast but lethal in chicken and human cells. T4p is associated with elongation and consistently roles are identified in RNA processing and 3' end processing. No specific kinase has been identified, although Plk1 (polo-like kinase) can catalyse T4p in human cells (Hintermair *et al.* 2012).

Each CTD repeat contains proline residues in positions 3 and 6 between the phosphorylation sites on the Rpb1-CTD. Proline residues can be isomerised to undergo conformational changes in *cis/trans* conformation by peptidyl-prolyl *cis-trans* isomerases (PPiases) (Egloff and Murphy 2008a; Dai *et al.* 2012). Pin1 (Ess1 in yeast) is known to specifically recognise the S(phospho)/T(phospho)–Proline motif and provide binding sites for the co-transcriptional machinery. Ssu72 (S5p phosphatase) preferentially dephosphorylates S5p in the Proline-6 in *cis*-conformation (Werner-Allen *et al.* 2011).

Glycosylation of serine and threonine residues can also occur on the Rpb1-CTD by addition of O-linked N-acetyl-glucosamine (O-GlcNAc)(Kelly *et al.* 1993). Phosphorylation and O-GlcNAcylation are mutually exclusive and the dynamic interplay is thought to be maintained by the concerted action of Ogt (O-GlcNAc transferase) and Oga (O-GlcNAc aminidase) proteins. Knockdown of Ogt proteins is also shown to decrease in transcription and RNAPII occupancy over B-cell promoters. The cycling of O-GlcNAc levels in higher eukaryotes is though to be important for specific gene transcription (Comer and Hart 2001; Ranuncolo *et al.* 2012).

The Rpb1-CTD is a target for ubiquitination and this modification is catalysed by protein Wwp2 (E3 ubiquitin ligase) in mouse stem cells (Li *et al.* 2007b). The lysine residues in the non-consensus repeats present in distal part of CTD are primary targets. Ubiquitination is known to occur at non-CTD residues and is believed to alter the subunit composition and complex assembly of RNAPII (Daulny *et al.* 2008).

Arginine methylation has also been more recently reported on the Rpb1-CTD on the single non-consensus Arginine residue at the proximal part in the non-consensus heptad repeat (Sims *et al.* 2011). Arginine methylation at repeat 31 (YSPSSP**R**) is catalysed by Carm1 protein (co-activator-associated arginine methylation) and this modification is implicated in inhibition of general expression of snRNA and snoRNAs (Chapman *et al.* 2008; Sims *et al.* 2011).

The CTD heptad repeat can be modified at all residues (phosphorylation, glycosylation, methylation, ubiquitination and isomerisation) leading to a combinatorial possibility of multiple CTD states. In addition, there are 52 repeats of the heptad in the mammalian CTD that further increases the permutation of combinations. New modifications and tandem combinations of modifications are increasingly being studied on the CTD that can further elucidate the functionality and role of the CTD code.

1.3. Active transcription cycle and CTD modifications

The transcription cycle at active protein-coding genes consists of recruitment of hypo-phosphorylated RNAPII onto the gene promoter, with the RNAPII undergoing transcription initiation, transition to elongation, elongation, termination and recycling, to produce a stable, mature mRNA that is exported to sites for translation and protein output of the encoded DNA message. The distinct phases of transcription are regulated, demarcated by specific modification on the Rpb1-CTD and further dynamic interactions with activators, repressor, chromatin remodellers, RNA processing machinery and co-transcriptional processing factors. The tight regulation of transcription and

its components keeps gene expression in control, allowing appropriate responses either to gene specific expression levels or global overhaul during differentiation or in response to environmental cues.

1.3.1. Initiation (PIC assembly and disassembly)

Transcription initiation is a complex, multistep process that involves the recruitment of hypo-phosphorylated RNA polymerase to gene promoters and accessibility to DNA around the transcription start site (TSS). Initiation is thought to be a rate-limiting step in the transcriptional process (Wade and Struhl 2008). RNAPII and basal transcription factors recognize and assemble on the promoters. The pre-initiation complex, corresponding to a minimal set of factors required for initiation, includes RNAPII, basal transcription factors (TFIIB, TFIID, TBP, TFIIE, TFIIH), and formation of the PIC requires dynamic interactions between RNAPII and basal factors to increase binding affinity to promoters and occupancy (Butler and Kadonaga 2002). Other factors include the Mediator complex, DNA binding co-activators, nucleosome remodelers and chromatin modifiers (Li *et al.* 2007a; Nechaev and Adelman 2011). The exact mechanism of Transcription Start Site (TSS) selection is thought to be specific to organismal diversity and chromatin architecture. Factors important for selection include positioning of RNAPII and GTF's to sequence specific factors, the presence of distinct sequence elements including a TATA box, Initiator, or Downstream Promoter Element (DPE), TFIIB recognition element and importantly CpG islands, amongst other genetic features.

Following PIC initiation, assembly and nucleosome repositioning, DNA unwinds and RNAPII initiates transcription. Promoter clearance and transition into elongation requires S5p on Rpb1-CTD by Cdk7 subunit of TFIIH and is also facilitated by promoter-proximal modified histones and other co-factors. Instability of the RNAPII complex during the early stages of transcription leads to abortive transcription. Following promoter escape and synthesis of an nascent transcript (~20 nucleotides), S5p recruits the capping enzyme and stabilises RNA by addition of the 7-methyl-guanosine cap. S5p also

coordinates and interacts with chromatin remodellers including Set family proteins. During the initial transcription steps, promoter proximal pausing has been demonstrated by stalling of transcriptionally engaged RNAPII (10-50 nucleotides) near the TSS by negative elongation factors.

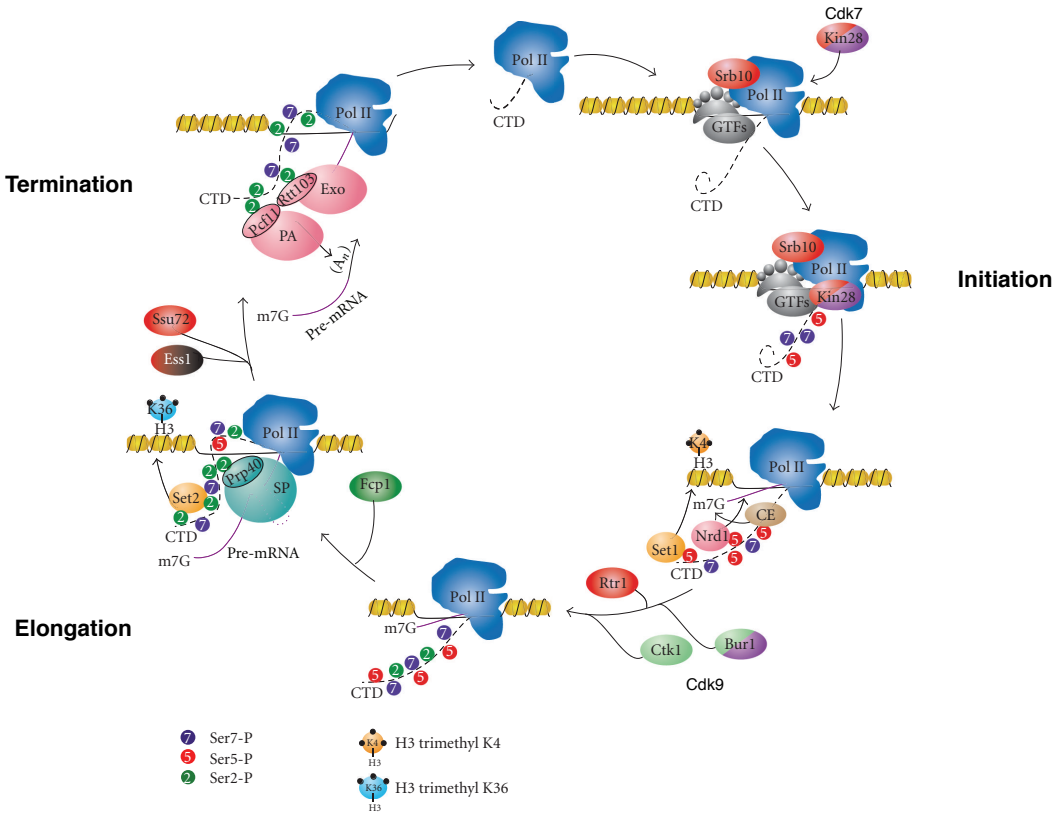


Figure 1.2 Transcription cycle at active genes and factors that interact with RNAPII during transcription. Recruitment of RNAPII at promoters, S5p and subsequently S7p on Rpb1-CTD associates with transcriptional initiation. S2p transitions RNAPII to elongation further interacting with range of protein cohorts to mature and stabilise RNAPII. Termination prepares RNAPII for next round of transcription. During the process of transcription, Rpb1-CTD interacts with range of chromatin remodellers, histone modifiers and components of transcription machinery as highlighted (Adapted from (Zhang *et al.* 2012b).

1.3.2. Elongation

Phosphorylation of S7p is thought as a transition from initiation to productive elongation and occupancy of S7p mirrors S5p. S7p precedes S2p for productive elongation and is important for transcription and 3′-end processing of snRNA genes (Komarnitsky *et al.* 2000). Although the exact function of S7p is still unclear, modification by Cdk7, enrichment at promoters and de-

phosphorylation by S5p phosphatase (Ssu72) suggests a general and important function in gene expression (Ghazy *et al.* 2009; Zhang *et al.* 2012a).

Transition into productive elongation is triggered by recruitment of the P-TEFb kinase (Cdk9) that phosphorylates the S2 residue on Rpb1-CTD and relieves the effect of negative elongation factors (including DSIF and NELF). S2p mediates interactions with set of diverse RNAPII-associated chromatin factors that favour progressive elongation by facilitating open chromatin architecture and RNA processing (Nechaev and Adelman 2011; Hsin and Manley 2012). S2p provides a platform for assembly of complexes that travel along with the RNAPII towards the end of the gene. These factors include splicing factors, RNA processing and chromatin remodellers (Phatnani and Greenleaf 2006). S2p levels increase through the gene body with lower S5p (than promoters) and once elongating, RNAPII is remarkably stable and transcribes kilobases without dissociation from the DNA template.

1.3.3. Termination and recycling of RNAPII

Transcription termination requires release of the RNA transcript from transcription complexes, canonical cleavage and processing by polyadenylation machinery to generate a nuclease-protective poly-A tail (Richard and Manley 2009). The downstream transcripts produced by RNAPII on DNA are chopped off by exonculease digestion and destabilisation of RNAPII is mediated by a range of chromatin factors and additional factors (CTD phosphatases) that directly associate with CTD and alter its conformation (Nechaev and Adelman 2011; Zhang *et al.* 2012b). Appropriate termination is essential for maturation of the RNA molecule. In addition, termination allows recycling of RNAPII to allow multiple rounds of transcription in rapid succession, thereby facilitating subsequent rounds of productive transcription (West *et al.* 2008).

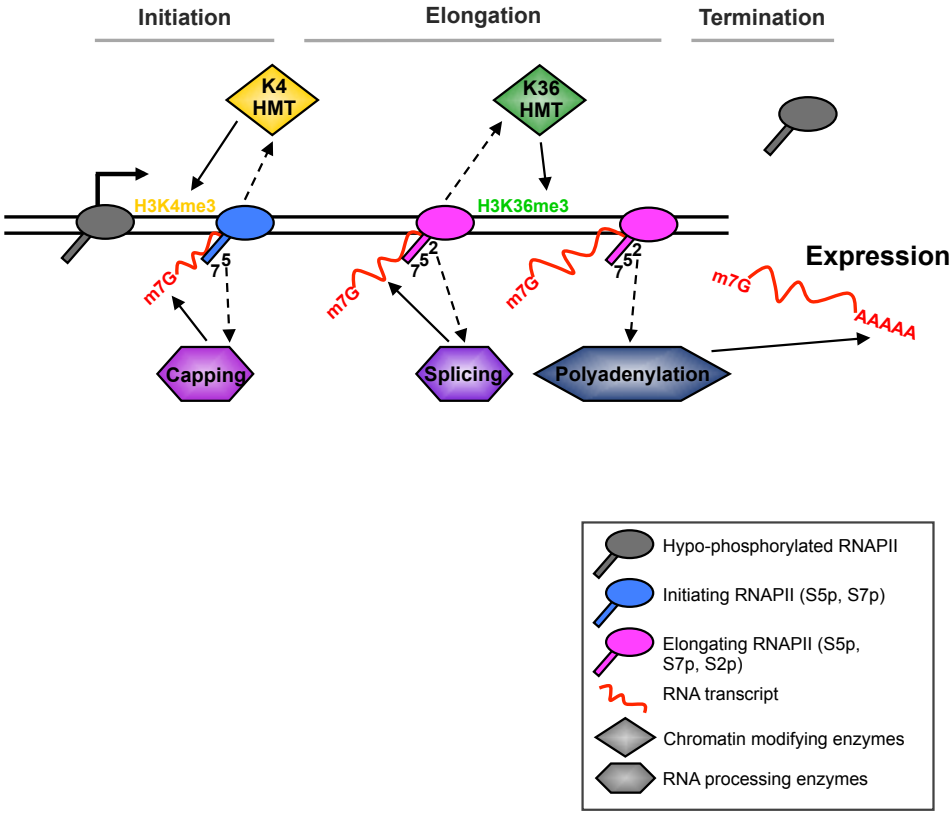


Figure 1.3 Transcription and coupling of co-transcriptional processes along with Rpb1-CTD modifications. (A) Initiation of transcription is coordinated with S5p, which further recruits capping machinery and associated histone and chromatin modifiers. RNAPII-S2p (S5p, S7p and S2p) is hallmark of transcription elongation that along with interaction with histone modifiers, also co-ordinates recruitment of co-transcriptional machinery, including splicing and polyadenylation machinery that processes the RNA to make a stable transcript. Termination requires the action of phosphatases on Rpb1-CTD to recycle and prepare hypo-phosphorylated RNAPII for the next round of transcription.

1.3.4. RNA polymerase II, protein interactions and stem cell specific co-factors

RNAPII and modifications to the Rpb1-CTD transcribe the genetic information into an mRNA message. The Rpb1-CTD acts as a platform whereby chromatin factors including remodellers, RNA processing and co-transcriptional machinery interact and regulate transcription. Dynamic modifications to the CTD transition RNAPII to different stages of transcription cycle along with recruitment of appropriate factors (Buratowski 2009).

During initiation, S5p associates with histone methyltransferases (Set1) leading to open chromatin confirmation marked by H3K4me3. In addition S5p

recruits capping enzyme to modify the 5' end of nascent RNAs (Fabrega *et al.* 2003). S5p physically associates with mRNA and capping enzyme to add a protective m7G (7-methyl guanosine) cap to the nascent RNA, while disrupting interactions with factors involved in formation of pre-initiation complex (mediators, basal transcription factors, etc.) (Proudfoot *et al.* 2002). S7p and S2p hallmarks of elongating RNAPII occur post-initiation (S5p) signalling in the transition to active elongation (Li *et al.* 2005). S7p and S2p integrate elongation with chromatin remodelling through the recruitment of Set2 (and H3K36me3) marking open chromatin structure compatible with elongation in gene bodies (Keogh *et al.* 2005). Transcription elongation also couples RNA synthesis with co-transcriptional RNA processing, splicing and polyadenylation factors (Kizer *et al.* 2005; Walsh *et al.* 2010). The polyadenylation factors and the termination factors preferentially associate with S2p and prepare RNA for maturation and protection while recycling RNAPII for next round of transcription.

Our knowledge of transcriptional process and components involved stems mostly from seminal work done in yeast. Quite a few yeast protein homologs exist in mammalian cells. However, the organismal complexity in mammals (e.g. 52 heptad repeats in the CTD including many non-canonical amino acids target for additional modifications) introduces overlapping function for homolog proteins and in addition often similar processes between yeast and mammalian are regulated by different cascades of protein cohorts. Stem cells additionally have specific sets of activators and co-activators that functionally interplay to maintain stem cell state (Fong *et al.* 2012). Remarkably, in stem cells a limited number of master transcription factors (Oct4, Sox2 and Nanog) define a stem cell-specific transcriptional signature by recruitment of specific factors including histone modifiers (e.g. p300/CBP, WDR5/Trithorax complex) and chromatin remodellers (esBAF) to ensure active chromatin architecture and RNAPII access (Ho *et al.* 2009; Ang *et al.* 2011; Fong *et al.* 2011). Additionally stem-cell specific transcriptional programs are mediated by

interactions with Mediator, TAFs/TFIID and the nucleotide excision repair (NER) complex (Fong *et al.* 2012).

TBP-associated factors (TAFs) are transcriptional co-activators that along with GTF's recognize promoters, anchor TFIID to nucleosomes leading to formation of a PIC (Pre-initiation complex). TAFs are known to have many paralogs, TBP-associated factors and importantly tissue specific expression. In ES cells, TAFs are highly expressed and down regulated in terminally differentiated cells (Deato and Tjian 2008; Goodrich and Tjian 2010). Genome-wide studies and proteomic studies have identified TAFs and TBP as important regulators of Oct4 in stem cells (Kagey *et al.* 2010; Ding *et al.* 2012). In addition, combinatorial assembly of cofactors (TAFs/TFIID, XPC/SCC, OCT4, SOX2) can led to transcriptional specificity in stem cells. TAFs are also important for pluripotency and Taf3 has been shown to mediate novel long-distance enhancer–promoter DNA interaction (with cohesion and mediator) in mouse ES cells (Liu *et al.* 2011).

The Mediator complex has recently been shown to interact with a wide array of transcriptional activators and RNAPII and coordinate stem cell specific functions (Kagey *et al.* 2010). Mediator complex functions as a docking site for RNAPII and TFIIH and facilitates promoter escape and transition to transcription elongation. Mediator in stem cells is required for proper expression of Oct4 and interactions with the cohesin complex stabilizes long-distance enhancer–promoter DNA interactions specific for ES cell genes. In stem cells, Mediator also functions as critical cofactor in a cascade of signalling pathways required for ES cellular maintenance and homeostasis (including Wnt signalling, BMP and TGFβ pathway) (Morris *et al.* 2008; Varelas *et al.* 2008). Other stem cell specific factors including chromatin remodellers (Chd1) (Gaspar-Maia *et al.* 2009), histone acetyltransferase (p300) (Black *et al.* 2006) and elongation factor (Paf1) (Ding *et al.* 2009) also play fundamental roles in coordinating with master regulators and regulating transcription.

1.4. The Polycomb group of proteins and silencing via chromatin modifications

The Polycomb Group (PcG) gene family is an evolutionarily conserved family of proteins, originally discovered in *D. melanogaster* as repressors of *Homeotic* genes that establish body plan and segmentation. Polycomb proteins are classified into two distinct multi-protein complexes names PRC1 (Polycomb repressive complex 1) and PRC2 (Polycomb repressive complex 2). PRC2 is involved in the initiation of silencing and is composed of three main subunits (Eed, Suz12 and Ezh2) containing histone de-acetylase and histone-methyltransferase activities. Ezh2 subunit can methylate histone H3K9me and H3K27me3, that mark silenced chromatin (also histone H1 lysine 26)(Cao *et al.* 2002; Valk-Lingbeek *et al.* 2004). PRC2 deletion is embryonic lethal and mutation in the catalytic domain has implications in development, pluripotency and differentiation. PRC1 is implicated in stable maintenance of gene repression, recognizes the PRC2-catalysed H3K27me3 and via its chromodomain catalyses H2Aub1. PRC1 complexes are diversified with the core complex being composed of Cbx proteins (Polycomb homolog), Ring1b, Ring1a while the other components inter-changeably form multiple PRC1 complexes. Both PRC2 and PRC1 interact with histone methyltransferases, histones and counteract SWI/SNF-chromatin-remodelling complexes (Breiling *et al.* 1999; Valk-Lingbeek *et al.* 2004).

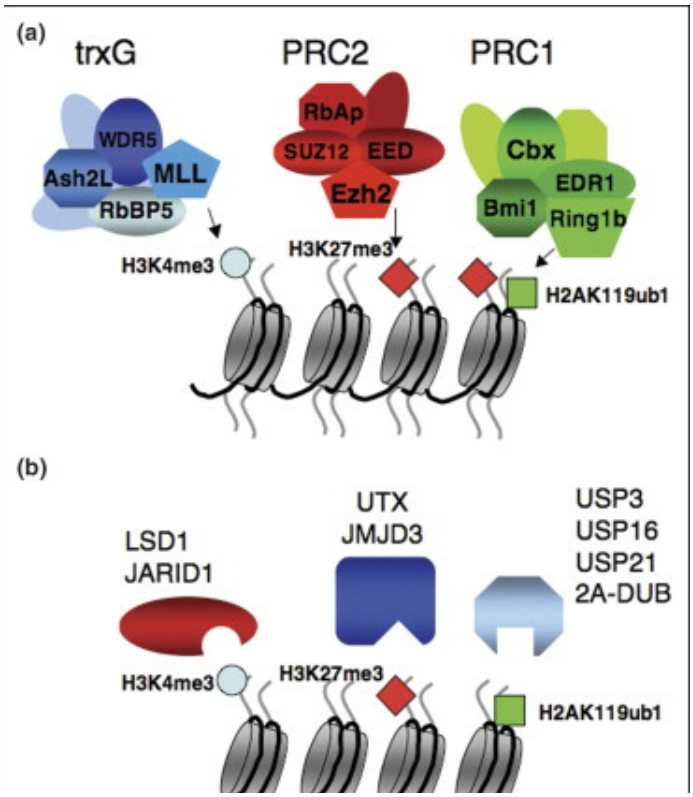


Figure 1.4 Polycomb proteins, their catalysed modifications and reversibility of Polycomb modifications. (a) Polycomb repressive complexes (PRC2, PRC1) catalyse repressive histone modifications (H3K27me3, H2Aub and H3K4me3) whereas Trithorax complex is associated with active histone marks (H3K4me3). (b) Both the active and repressive histone modifications are reversible by action of specific demethylases, de-ubiquitinating enzymes and ubiquitin-specific proteases (from (Lund and van Lohuizen 2004).

Table 1.2 Polycomb repressive complex proteins and their mouse counterparts (from (Lund and van Lohuizen 2004).

Drosophila proteins	Mouse proteins
PRC2/Initiation complex	
Esc (Extra sex combs)	Eed
E(z) (Enhancer of Zeste)	Ezh1/Enx2
	Ezh2/Enx1
Su(z)12 (Suppressor of Zeste 12)	Suz12
PRC1/Maintenance complex	
Pc (Polycomb)	Cbx2/M33
	Cbx4/Mpc2
Ph (Polyhomeotic)	Edr1/Mph1/Rae28
dRING (Really Interesting New Gene)	Ring1/Ring1a
	Rnf2/Ring1b
Psc (Posterior sex combs)	Bmi1
	Rnf110/Zfp144/Mel-18
	Znf134/Mblr

Pho (Pleiohomeotic)	Yy1
Scm (Sex combs on midleg)	Scmh1
	Scmh2

1.4.1. Polycomb proteins and stem cells

Stem cells possess the unique capacity of self-renewal and the ability to differentiate into all lineages. Polycomb proteins have essential roles in embryonic development and consistent with this, loss of Polycomb proteins has implications for pluripotency, differentiation potential and even viability (O'Carroll *et al.* 2001; Valk-Lingbeek *et al.* 2004). Knockout of Polycomb proteins in ES cells causes de-repression of important developmental regulator genes and subsequent ES cell differentiation (Azuara *et al.* 2006; Stock *et al.* 2007b; Brookes *et al.* 2012).

1.4.2. Polycomb proteins and transcription regulation

The association of Polycomb proteins together with RNAPII, TFs and other chromatin proteins suggests a complex mechanism of PRC regulation along with transcriptional machinery downstream of transcription initiation (Brookes and Pombo 2009a).

Polycomb repression in *D. melanogaster* is mediated by short sequence motifs called PREs (Polycomb response elements). The PREs act as sequence signals for binding of Polycomb proteins and their subsequent repression. In *D. melanogaster*, insertion of a PRE within an active gene leads to onset of transcription by RNAPII but PREs prevent appropriate expression (Dellino *et al.* 2004; Schwartz *et al.* 2004). In mES cells, RNAPII-S5p complexes bind at the promoters of PRC-repressed genes and transition into coding regions. However, in the absence of S2p (Stock *et al.* 2007b; Brookes *et al.* 2012) loss of PRC1 components results in de-repression at PRC-repressed genes, suggesting the mechanistic link between PRC1 and Rpb1-CTD and not with elongation complexes (Brookes and Pombo 2009a). PREs or PRE-like mechanisms have not yet been identified in mES cells, suggesting a more complex recruitment and dynamic interplay for transcriptional repression. A combination of several TFs, including master

regulators (Oct4, Nanog, Sox2), CTCF, E2F1 and Myc are known to associate with Polycomb proteins in mES cells. In addition, there is growing evidence that short and long ncRNAs mediate Polycomb recruitment to target sites.

1.4.3. Bivalency and poised genes in mES cells

In mES cells, important developmental regulator genes are bound by Polycomb proteins and harbour bivalent chromatin modifications that are thought to keep these genes poised for future activation (Azuara *et al.* 2006; Bernstein *et al.* 2006; Mikkelsen *et al.* 2007). The chromatin at PRC-repressed genes contains Polycomb-mediated H3K27me3, H2Aub1 along with active modification H3K4me3 (Brookes and Pombo 2009a). These bivalent domains have been identified in human ES cells, mES cells and also in differentiated cells, albeit to a much lower extent, that as in differentiated cells they resolve into monovalent configurations (Mikkelsen *et al.* 2007; Pan *et al.* 2007). The PRC-repressed genes are enriched for transcription factors that are critical for differentiation and development, and the bivalent chromatin configuration *i.e.* repressive histone modifications silence their expression in mES cells for pluripotent characteristics while the active chromatin marks allow for later activation (Brookes and Pombo 2009a).

Both PRC2 and PRC1 components occupy PRC-repressed genes along with their respective histone modifications H3K27me3 and H2Aub1 along with cohorts of transcription factors and chromatin co-factors (Stock *et al.* 2007b). Despite PRC-mediated repression, RNAPII is detected at promoters and coding regions of PRC-repressed genes in an unusual conformation incompatible with gene expression (Stock *et al.* 2007b). At these genes, RNAPII-S5p (initiating) is detected at promoters and coding regions in the absence of S7p or S2p (elongation). In addition, low level transcripts are identified at these genes. Interestingly, removal of Polycomb protein Ring1b cause de-repression of these genes although without characteristic transcription elongation (mediated by RNAPII-S2p), suggesting a dynamic

interplay between Rpb1-CTD and Polycomb proteins (Stock *et al.* 2007b; Brookes *et al.* 2012).

1.4.4. Transcription cycle at Polycomb repressed genes and CTD modifications

The dynamics of gene expression regulation at PRC-repressed genes is quite complex. At PRC-repressed genes, Polycomb proteins, their histone modifications, active histone modifications and unusual RNAPII conformation mark and regulate their chromatin environment. In stem cells, this regulation is further kept in check by important stem cell master regulator transcription factors (Oct4, Sox2 and Nanog) that regulate the gene expression maintaining pluripotency and the capacity to self-renew.

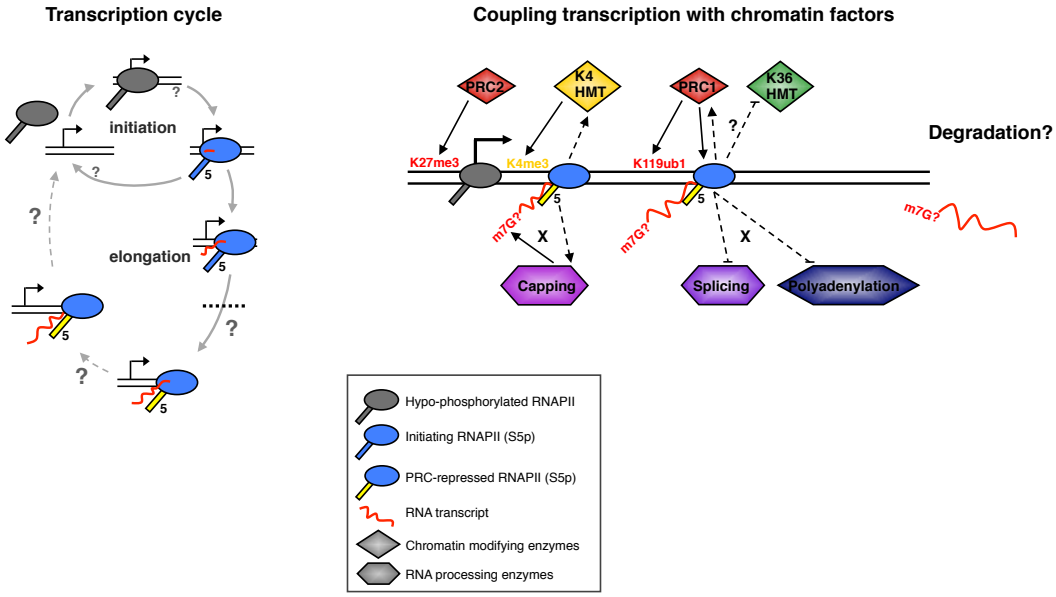


Figure 1.5 Transcription cycle at PRC-repressed genes and chromatin assembly. (A) RNAPII is present in an unusual state at PRC-repressed genes marked by only S5p and present at promoter and coding regions of genes. (B) Polycomb proteins mark repressive histone modifications and the presence of H3K4me3 (active mark) mediates a bivalent chromatin architecture and novel RNAPII binding pattern at these genes. No mRNA is produced at PRC-repressed genes, highlighting the absence of S2p and co-transcriptional processing machinery (adapted from (Brookes and Pombo 2009b).

Genome wide analysis by ChIP-Sequencing (ChIP-Seq) from our group has unravelled that 30% of Refseq genes constitute PRC-repressed genes. The

characteristic hallmark of PRC-repressed genes is an unusual RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) and histone modifications (H3K4me3, H3K27me3, H2Aub). Analysis by single ChIP and sequential-ChIP confirms the co-existence of both RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) and Polycomb proteins (Ezh2 and Ring1b) at these genes. RNAPII-S5p at these genes extends from the promoters to the coding region of the genes without any S7p or S2p or mRNA production, consistent with impaired elongation and lack of association with the RNA processing machinery. Conditional knockout of the Polycomb protein Ring1b leads to rapid de-repression without characteristic S2p elongation further highlighting the role of Rpb1-CTD in the mechanistic regulation at these genes.

From the genome-wide analysis in mES cells, broadly three groups of genes were observed (and sub-groups). The first group of genes are actively transcribing genes that contained characteristic active RNAPII configuration ($S5p^{+}S7p^{+}S2p^{+}$), without any repressive histone modification or Polycomb occupancy. The next group of genes includes Polycomb-repressed genes that contained RNAPII ($S5p^{+}S7p^{-}S2p^{-}$), Polycomb proteins (Ring1b and Ezh2) and repressive histone modifications (H3K27me3 and H2Aub1) (Brookes *et al.* 2012). PRC-repressed genes were further classified based on levels of occupancy of RNAPII and Polycomb proteins and 4 broad sub-classes were identified in a gradient range of Polycomb regulation *i.e*. PRC-active, PRC-intermediate, PRC-repressed and PRC-only. The last group of genes consist of inactive genes that do not harbour either RNAPII or Polycomb proteins (Brookes *et al.* 2012). Genome-wide classification of PRC-repressed genes has revealed several important biological processes these genes encode for, inlcuding important developmental regulators genes, metabolic proteins (Pyruvate metabolism, TCA cycle, Glycolysis, Notch signalling etc.), signalling pathways (TGFβ, MAPK, Wnt, p53 signalling pathways) (Brookes *et al.* 2012).

1.5. Proteomics and Mass spectrometry

The global analysis of proteins, which are the key functional entities in the cell, arguably forms the principal level of information required to understand how cells function; such analysis is referred to as proteomics (Altelaar *et al.* 2013). The different disciplines that contribute to proteomics include cell imaging (Light and electron microscopy), Microarray, ChIP experiments, and genetic readout experiments, as exemplified by the yeast two-hybrid assay (Aebersold and Mann 2003).

In recent years, proteomics technologies particularly mass spectrometry (MS)-based protein identification has matured immensely and has been applied to gain significant biological insights. Various MS-based proteomics approaches have been developed and applied to various biological systems including label-free protein quantification, label-based protein quantification, proteome expression profiling, characterization of proteomic repertoires, comparative gene expression profiling, absolute protein quantification and characterisation of dynamic and spatial protein distribution (Ong and Mann 2005; Altelaar *et al.* 2013). More recently several clinical applications have yielded significant insights and these include integrative omics profiling to identifu molecular causes of disease, identification of clinical biomarkers, understanding cellular heterogeneity and personalized cancer therapies (Zhou *et al.* 2012; Altelaar *et al.* 2013).

Stable isotope labelling of amino acids in cell culture (SILAC) offers a simple and practical approach to perform quantitative proteomics (Ong *et al.* 2002). SILAC labelling methods have been further developed and applied to generate entire fly and mouse (Sury *et al.* 2010; Zanivan *et al.* 2012). Additionally SILAC approaches have also been applied to distinguish proteomic differences between control and disease samples (Mann 2006; Sury *et al.* 2010; Zanivan *et al.* 2012).

The dramatic and ongoing improvements in MS technologies have accelerated the application of MS-based proteomics to detective sensitively and robustly proteins, their abundance and dynamic regulatory network in different biological applications. The field of proteomics is geared towards unravelling novel proteins interactions, structural dynamics, coupling imaging with MS (Imaging MS), whole proteome identifications and genome-based proteome (Proteo-genomics). Recent advancements in single cell proteomics have further highlighted the potential of MS based approaches in developing our understanding of biological processes.

1.6. Research aims and objectives

Understanding the regulation of RNAPII and Polycomb interplay in stem cells provides significant insights into stem cell pluripotency, dynamic chromatin architecture and gene regulation in mES cells and during lineage specification. These aspects are critical to understand development and developmental regulation of an organism. To understand the dynamic interplay and dissect the components involved, I aimed to investigate the RNAPII-Polycomb interplay in mES cells from a proteomic point of view. I developed an unbiased method called ‘Proteome-ChIP’ (pChIP) that unravels the chromatin bound proteome. To identify the chromatin landscape associated with RNAPII, I applied pChIP to qualitative and quantitatively dissect RNAPII modifications and highlight their proteome composition. Towards understanding the complex RNAPII regulation in mES cells and identifying the plethora of RNAPII modifications and their mechanistic processes, I applied a Systems Biology approach to comprehensively dissect RNAPII proteome and unravel protein associations and their dependencies on RNAPII modifications (S5p, S7p and S2p). The analysis also serves to provide an invaluable resource of RNAPII modifications and respective protein dependencies.

Further extending our analysis and challenging the current technologies, I also aimed to perform pChIP on lower complexity samples (Native-ChIP and

Gradient-ChIP) to unravel the chromatin landscape. The work in this thesis focuses on characterisation of RNAPII-bound chromatin proteome in mES cells and unravelling of novel RNAPII processes specific for ES cells and with implications for stem cell biology.

1.7. Thesis outline

The thesis starts with literature review (Chapter 1) of mechanism of epigenetic modifications and gene regulation on chromatin, particularly focusing on RNAPII regulation, transcriptional regulation and RNAPII-Polycomb interplay in stem cells. Chapter 2 describes the briefly the experimental methods and approaches used for this thesis. Results are described in Chapter 3 to 8. In Chapter 3, I highlight the mES cells chromatin complexity, optimise and develop the pChIP method along with RNAPII-pChIP as proof of principle. In Chapter 4, I optimise and extend pChIP to quantitatively explore the RNAPII proteome using SILAC methodology. Chapter 5 describes the comprehensive pChIP experimental setup and using a simple classification to unravel the chromatin landscape and dependencies on RNAPII modifications. I also uncover and discuss novel S5p only processes previously not identified. In Chapter 6, I employ a Systems Biology approach (in collaboration) to dissect and unravel patterns within the pChIP dataset and uncover novel associations not previously identified by conventional approaches. Chapter 7 includes bioinformatic comparisons with published datasets for mRNA-bound proteome and RNAPII interactome in mitosis. Finally, in Chapter 8, I extend RNAPII-pChIP to native chromatin and perform gradient fractionation to obtain purified crosslinked chromatin sample for RNAPII-pChIP. Lastly and importantly, I discuss the results and their implication with future research directions in Chapter 9.

2. Materials and Methods

2.1. Murine embryonic stem cell culture

2.1.1. Murine ES-OS25 cell culture

Mouse ES-OS25 cells (kindly donated by A. Smith) were grown on 0.1% gelatin-coated surfaces in GMEM-BHK21 supplemented with 10% fetal calf serum (FCS), 2 mM L-glutamine, 1% MEM non-essential amino acids (NEAA), 1 mM sodium pyruvate, 50 µM 2-mercaptoethanol, 1000 U/ml of human recombinant leukaemia inhibitory factor (LIF; Chemicon, Millipore, Chandler's Ford, UK) and 0.1 mg/ml Hygromycin (Roche) as described previously (Niwa *et al.* 2000; Billon *et al.* 2002).

2.1.2. Murine ES-OS25 SILAC cell culture

Mouse ES-OS25 cells were grown for a minimum of 4 passages before SILAC labelling. SILAC cells were grown on 0.1% gelatin-coated surface with SILAC-DMEM (without lysine and arginine amino acids) supplemented with 15% knockout serum replacement (KOSR), 2 mM L-glutamine, 1% MEM non-essential amino acids (NEAA), 1 mM sodium pyruvate, 50 µM 2-mercaptoethanol, 1000 U/ml of human recombinant leukaemia inhibitory factor (LIF) and 0.1 mg/ml Hygromycin (Roche). For SILAC-light cells, L-lysine-HCl (($^{12}C_6\,H_{14}\,^{14}N_2\,O_2$); 0.8mM) and L-arginine-HCl ($^{12}C_6\,H_{14}\,^{14}N_4\,O_2$; 0.4mM) were added to media. For SILAC-heavy cells, L-lysine-HCl ($^{13}C_6\,H_{14}\,^{15}N_2\,O_2$; 0.798mM) and L-arginine-HCl ($^{13}C_6\,H_{14}\,^{15}N_4\,O_2$; 0.398mM) were added to media.

2.2. DNA-Chromatin immunoprecipitation

Antibodies used for chromatin immunoprecipitation (ChIP), including control and bridging antibodies, are presented in Table 2.1. Antibodies towards different epitopes of RPB1 are schematically represented in Fig. 2.1.

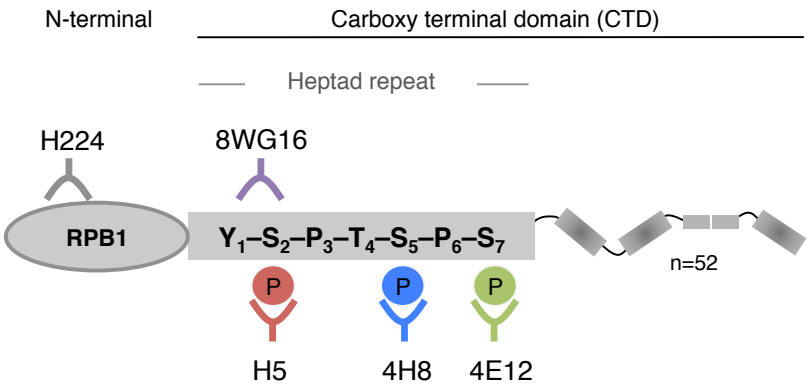


Figure 2.1 RPB1 antibodies recognising specific phospho-epitopes. H224 binds outside of the RPB1 CTD and so recognises both hypo- and hyper-phosphorylated RPB1. Clone 4H8 recognises phosphorylated serine residues (position 5; S5p; blue). Clones H5, 4E12 and 3D12 are used for detection of S2p (red) and S7p (green) respectively. 8WG16 (purple) recognises un-phosphorylated S2 residues.

2.2.1. Fixed chromatin preparation

2.2.2. Fixed chromatin preparation (Formaldehyde crosslinked)

Fixed chromatin was prepared as described previously (Stock *et al.* 2007a). Briefly, cells were treated with 1% formaldehyde (37°C, 10 min) and the reaction was stopped by addition of glycine to a final concentration of 0.125 M. Cells were washed in ice-cold PBS, before “swelling” buffer (25 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl and 0.1% NP40) was added to lyse the cells (4°C, 10 min). Cells were scraped from flasks, and nuclei isolated by Dounce homogenization (60 strokes, “Tight” pestle) and centrifugation. After re-suspension in “sonication” buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X100, 0.1% Na-deoxycholate and 0.1% SDS), nuclei were sonicated to produce DNA fragments with a average length of <1.6 kb (Stock *et al.* 2007a) using a Diagenode Bioruptor (Liege, Belgium; full power; 1h: 30s ‘on’, 30s ‘off’; 4°C) or using a Diagenode Bioruptor-Plus (full power; 30 cycles: 30s ‘on’, 30s ‘off’; 4°C). The resulting material was centrifuged twice (4°C, 10 min) at 14,000 rpm to remove insoluble material. Swelling and sonication buffers were supplemented with phosphatase inhibitors 5 mM NaF, 2 mM Na_3VO_4, 1 mM PMSF, and protease inhibitor cocktail (Roche, Burgess Hill, UK).

"Chromatin concentration" was estimated by measuring absorbance (280 nm) of alkaline-lysed, crosslinked chromatin, and converting into arbitrary chromatin units using the conversion 50 mg/ml for 1 absorbance unit.

2.2.3. Fixed chromatin preparation (Double crosslinked – EGS and Formaldehyde)

Fixed chromatin was prepared as described previously (Stock *et al.* 2007a). Briefly, cells were treated with 1.5mM EGS (*Ethylene glycolbis [succinimidyl succinate]* for 37°C, 10 min) and subsequently with 1% formaldehyde (37°C, 10 min) and the reaction was stopped by addition of glycine to a final concentration of 0.125 M. Rest of the protocol was same as that formaldehyde crosslinked ChIP.

2.2.4. Native chromatin preparation

Preparation of unfixed chromatin was carried out as described previously (O'Neill and Turner 2003; Szutorisz *et al.* 2005) with some modifications, and used for histone modification and RNAPII modification ChIP. Cells were washed in ice-cold PBS and incubated in 0.5% NP40 in 1X TBS (0.01 M Tris-HCl pH 7.4, 3 mM $CaCl_2$, 2 mM $MgCl_2$) for 10 min. Cells were scraped and incubated on ice (50 min) to complete cell lysis. Nuclei were extruded by Dounce homogenization (60 strokes, "Tight" pestle) and re-suspended in 25% sucrose in 1X TBS. 50% sucrose in 1X TBS was used to underlie this suspension. The nuclear pellet was washed once in 25% sucrose in 1X TBS and then re-suspended in "digestion" buffer (50 mM Tris-HCl pH 7.4, 1 mM $CaCl_2$, 4 mM $MgCl_2$, 0.32 M sucrose) to a concentration of 0.5 mg DNA/ml (determined by measuring absorbance of alkaline-lysed chromatin).

Chromatin was digested with 2U/ml micrococcal nuclease (MN, Sigma) to produce fragments containing mainly mono- and di-nucleosomes (37°C, 10 min; for optimisation of MN digestion see Fig. 5.9). Digestion was stopped by addition of 5 mM EDTA on ice. The first supernatant (S1) was recovered, and the pellet re-suspended in "lysis" buffer (1 mM Tris-HCl pH 7.4, 0.2 mM EDTA). After nuclear lysis was completed (30 min on ice and overnight at -

20°C) a second supernatant (S2) was collected. After fragment size analysis, supernatants (S1 and S2) were combined. TBS, digestion and lysis buffers were supplemented with 5 mM NaF, 2 mM Na_3VO_4, 1 mM PMSF, and protease inhibitor cocktail (Roche, Burgess Hill, UK).

"Chromatin concentration" was obtained by measuring absorbance (280 nm) of alkaline-lysed, native chromatin, and converting into arbitrary chromatin units using the conversion 50 mg/ml for 1 absorbance unit.

2.2.5. Fixed DNA-ChIP with cesium chloride gradient (gradient-ChIP)

Fixed chromatin (formaldehyde crosslinked) was prepared as described (Section 2.2.1.1). Chromatin was re-suspended in cesium chloride solution ($CsCl_2$; Final concentration- 567.8mg/ml) and ultra-centrifuged (Sw55Ti, Beckmann coulter, 40,000 rpm, 72 hrs, 4°C). After gradient separation, 10-11 fractions were collected (using low flow rate) and excess salt removed by dialysis with sonication buffer. Fractions 1-4 were pooled together to form 'DNA-only' fraction, fraction 5-8 were pooled to form 'nucleo-histone' fraction and fraction 9-10 were pooled to form 'protein-only' fractions.

DNA concentration was measured from the 3 fractions (DNA-only fraction, nucleo-histone fraction and protein-only) as described (Section 2.2.7) and corresponding chromatin concentration was used for pChIP.

2.2.6. Confirmation of fragment size of native or fixed chromatin

To confirm appropriate shearing or enzymatic fragmentation had occurred, DNA was purified from chromatin and subjected to Agarose electrophoreses. Fixed chromatin first had aldehyde cross-links reversed by adding NaCl and RNase A to final concentrations of 160 mM and 20 µg/ml, respectively, and incubating at 65°C overnight. With all samples, the EDTA concentration was adjusted to 5 mM, and then 200 µg/ml proteinase K (Roche) was added (45°C, 2h). DNA was recovered by phenol-chloroform extraction and ethanol precipitation, and re-suspended in TE. DNA was separated on a 1% Agarose gel to check fragmentation efficiency.

2.2.7. Immunoprecipitation with magnetic beads

For mouse IgG and IgG antibodies, protein-G-magnetic beads (Active Motif) were incubated with rabbit anti-mouse (IgG+IgM) or anti-IgM bridging antibodies, respectively (Jackson Immunoresearch; 10 µg per 50 µl beads) for 1h (4ºC) and washed with sonication buffer. For rabbit antibodies, magnetic beads were just washed with sonication buffer. Fixed (700 µg) or native chromatin (250 µg) was immunoprecipitated (4ºC, overnight) with 10-50 µg antibody and 50 µl beads (with/without bridging antibody).

2.2.8. ChIP washes and elution's

For Protein A/G immunoprecipitations, beads were washed (1x) with sonication buffer, (1x) sonication buffer containing 500 mM NaCl, (1x) 20 mM Tris pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP40 and 0.5% Na deoxycholate, and (2x) TE buffer (1 mM EDTA, 10 mM Tris HCl pH 8.0).

After immunoprecipitations using IgM antibodies, beads were washed (2x) with sonication buffer, (1x) 2 mM Tris pH 8.0, 0.02 mM EDTA, 50 mM LiCl, 0.1% NP40 and 0.1% Na-deoxycholate; and (1x) TE buffer. Immune complexes were eluted from beads (65°C, 5 min; and room temperature, 15 min) with 50 mM Tris pH 8.0, 1 mM EDTA and 1% SDS. The elution was repeated and eluates pooled.

2.2.9. DNA purification, quantification and analysis

For fixed chromatin samples, reverse cross-linking was carried out (16h, 65°C) with addition of NaCl and RNase A to final concentrations 160 mM and 20 µg/ml, respectively. Native and fixed samples had EDTA increased to a final concentration of 5 mM and samples were incubated with 200 µg/ml proteinase K (50°C, 2h). DNA was recovered by phenol-chloroform extraction and ethanol precipitation. The final DNA concentration was determined by PicoGreen fluorimetry (Molecular Probes, Invitrogen) and samples were diluted to the same concentration (0.2 ng/µl). The same amount (0.5 ng) of immunoprecipitated and input DNA were analysed by quantitative real-time PCR (RT-PCR). Amplifications (40 cycles) were performed using SensiMix

NoRef (Quantace, London, UK) with DNA Engine Opticon 1/2 RT-PCR system (BioRad, Hemel Hempstead, Hertfordshire, UK).

IP or control “cycle over threshold” (Ct) values from the quantitative PCR (IP) were subtracted from the input Ct values (Input Ct). This figure was converted into the fold enrichment by $2^{(\text{input Ct} - \text{IP})}$. Histone modifications levels were normalized to levels of the core histone in some situations.

2.3. Protein-chromatin immunoprecipitation

2.3.1. Fixed pChIP, reverse crosslinking and protein elution

Fixed chromatin was prepared as described (Section 2.2.1.1) and gradient fractions were obtained as described (Section 2.2.3). pChIP’s were set up using magnetic beads and bridging antibody (Section 2.2.5). ChIP washes were performed as normal (Section 2.2.6). After the final wash in TE buffer, fixed immunoprecipitated chromatin complexes were treated with Benzonase (25 units) in 30µl DNase buffer at 37°C, 30 min; Merck) to chop DNA fragments and to allow efficient elution of proteins. Proteins were reverse crosslinking and eluted from the beads twice (first at 60°C, O/N min; and second time at 95°C, 15 min) using 30 µl of custom Laemmli buffer (130 mM Tris-HCl pH 6.8, 20% glycerol, 200mM DTT, 0.02% bromophenol blue, and 4% SDS) per pChIP elution, giving a final eluate volume of 60 µl.

Unlike DNA, the proteins in the samples cannot be amplified by a PCR and it is very important to elute as much as possible from beads and also to make sure that no remaining beads are present in sample. Any bead contaminants will clog the SDS-PAGE gel, not allowing the proper separation of proteins. We recommend spinning the eluted 60µl samples twice after elution, recovering a cleared supernatant to completely remove any traces of beads, which is then frozen. As repeat freeze-thaw cycles can degrade the proteins in buffer, we recommend running the SDS-PAGE on the same day as protein elution.

Specificity of pChIP is also critical and we recommend comparison with No antibody control and a non-specific antibody (we use mouse anti-Digoxigenin) to filter non-specific binders from common contaminants. Robustness and specific enrichment of pChIP should be measured by running the input chromatin sample along with pChIP in Mass Spectrometry (MS).

2.3.2. Native pChIP, reverse crosslinking and protein elution

Native chromatin was prepared as described (Section 2.2.2) and the pChIP's were set up using magnetic beads and bridging antibody (Section 2.2.5). ChIP washes were performed as normal (Section 2.2.6). After the final wash in TE buffer, native immunoprecipitated chromatin complexes were reverse crosslinked and eluted from the beads twice (95°C, 10 min), using 60 µl of custom Laemmli buffer (130 mM Tris pH 6.8, 20% glycerol, 200mM DTT, 0.02% bromophenol blue, and 4% SDS) per pChIP elution.

Native protein complexes are more susceptible to degradation, so we recommend to use the samples straightaway and avoid long-term storage. As with fixed pChIP, care should be taken to remove any bead contamination in sample.

2.3.3. Fixed pChIP with cesium chloride gradient (modified pChIP)

Fixed chromatin was prepared as described (Section 2.2.2) and the pChIP's were set up using magnetic beads and bridging antibody (Section 2.2.5). ChIP washes were performed as normal (Section 2.2.6). After the final wash in TE buffer, native immunoprecipitated chromatin complexes were reverse crosslinking and eluted from the beads twice (95°C, 10 min), using 60 µl of custom Laemmli buffer (130 mM Tris pH 6.8, 20% Glycerol, 200mM DTT, 0.02% bromophenol blue, and 4% SDS) per pChIP elution.

Native protein complexes are more susceptible to degradation and we recommend using the samples straightaway and avoid long-term storage. As with fixed pChIP, care should be taken to remove any bead contamination in sample.

2.4. Western analysis

Antibodies used for western analysis, including loading control antibodies, are presented in Table 2.2. Antibodies towards different epitopes of RPB1 are schematically represented in Fig. 2.1.

2.4.1. RNAPII western analysis

Whole cell extracts for RNAPII westerns were prepared by lysing cells in ice-cold “lysis” buffer (Daniel and Carling 2002), scrapping, and shearing DNA by passage through a 25G needle. Total protein concentration was determined by Bradford assay. Cell lysates (0.5 μg total protein for 4H8 antibody, 10 μg for 4E12, 3E8 and Y1P antibodies, 5 μg for all other RNAPII antibodies) were resolved on 7.5% Tris-HCl or 3-8% Tris-acetate SDS-PAGE gels. Chromatin and pChIP samples for RNAPII westerns were resolved in a 10% Tris-HCl or 3-8% Tris-acetate SDS-PAGE gels.

Membranes were blocked (1h), incubated (2h) with primary antibody, washed, and incubated (1h) with HRP-conjugated secondary antibodies, all in blocking buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% Tween-20,; 5% non-fat dry milk). Membranes were washed (30 min) in blocking buffer without milk and briefly in 0.1% Tween-20 in PBS. HRP-conjugated antibodies were detected with ECL western blotting detection reagents (Amersham), according to the manufacturer’s instructions.

2.4.2. Alkaline phosphatase treatment

To dephosphorylate RPB1 after electrophoresis and transfer, membranes were incubated (37°C, 1h) in 0.1 U/μl alkaline phosphatase (AP) in NEB buffer 3 (New England Biolabs, Hitchin, UK) prior to blocking.

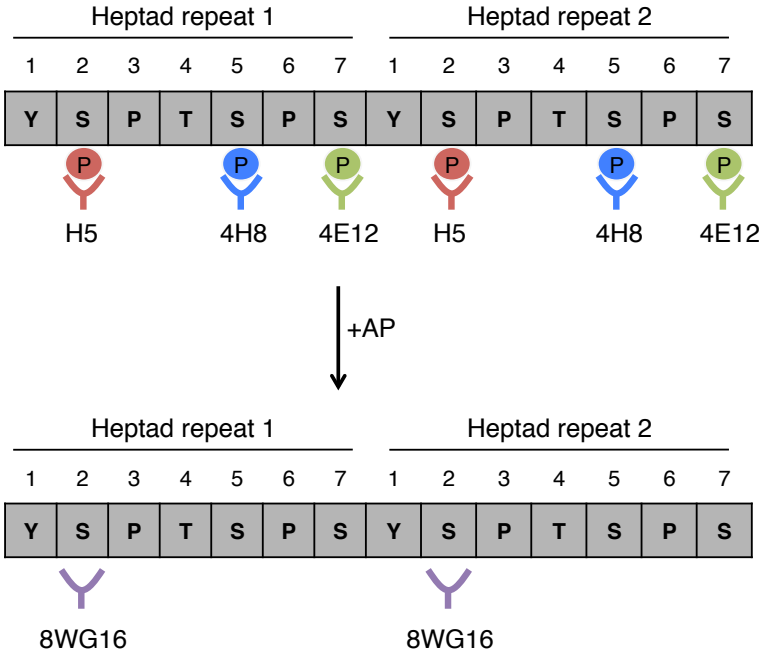


Figure 2.2 Alkaline phosphatase treatment de-phosphorylates Rpb1. Alkaline phosphatase (AP) removes phosphates from S5 (blue), S2 (pink) and S7 (green) residues, and allows reactivity with 8WG16 (purple). Antibody against Rpb1 N-terminal (H224) is not affected by AP treatment.

2.4.3. Polycomb, histone and other western analysis

Nuclear and histone extracts were prepared as described previously (Kuzmichev *et al.* 2002; de Napoles *et al.* 2004). Nuclear extracts (20 μg) or histone extracts (5 μg) were loaded on 10-15% gradient SDS-PAGE gels.

2.5. Imaging

2.5.1. Whole cell immunofluorescence for S5p and Nanog

Immunofluorescence was carried out essentially as described previously (Xie *et al.* 2006). Cells were cultured on sterile coverslips (No. 1.5, Agar Scientific) to achieve ~70% confluence on the day of fixation. Coverslips were fixed in 4% PFA, 125 mM HEPES, 0.1% Triton X100 (4°C, 10 min) and in 8% PFA in 125 mM HEPES (4°C, 30 min; (Guillot *et al.* 2004). Cells were permeabilized in 0.5% Triton X100 in PBS (60 min), and then incubated with 20 mM glycine in PBS (30 min). Cells were blocked (1h), incubated with respective antibodies (3h), washed (1.5h), incubated (1h) with FITC-conjugated donkey anti-mouse Ig (1:100; Jackson Immunoresearch Laboratories), and washed (4°C,

overnight); all in PBS+ (PBS supplemented with 0.1% casein, 1% BSA, 0.2% fish skin gelatin, pH 7.8). Coverslips were washed in 0.1% Tween-20 in PBS, incubated with TOTO-3 in 0.1% Tween-20 in PBS (1:500, 15 min), and washed in 0.1% Tween-20 in PBS and then PBS, before coverslips were mounted in VectaShield (Vector Labs), immediately before imaging.

2.5.2. Microscopy

Images were collected sequentially on a Leica SP2 confocal laser-scanning microscope (100x objective; NA 1.3) using Green Argon (488 nm) and Far Red (633 nm) lasers and pinhole equivalent to 1 Airy disk. For comparison of different treatments, images were collected on the same day using the same settings, and without saturation of the intensity signal. Images were transferred to Adobe Photoshop and contrast stretched with the same settings for all conditions.

Table 2.1 Antibodies used for ChIP analysis

Antibody against	Clone	Raised in (isotype)	Volume used in ChIP (per IP) Magnetic beads	Origin
RNAPII				
RNAPII S2p	H5 (MMS-129R)	Mouse (IgM)	20 µl	Covance, Princeton, NJ
RNAPII S5p	CTD4H8 (MMS-128P)	Mouse (IgG)	10 µl (25 µg)	Covance
RNAPII S7p	4E12	Rat (IgG, hybridoma)	100 µl	Kind gift from Dirk Eick
Non-phosphorylated S2 CTD residues	8WG16 (MMS-126R)	Mouse (IgG)	10 µl (25 µg)	Covance
RPB1 N-terminus (amino acids 1-224)	H224 (sc-9001x)	Rabbit (IgG)	-	Santa Cruz Biotechnology, Santa Cruz, CA
Histones and modifications				
H2Aub1	E6C5 (05-678)	Mouse (IgM)	50 µl	Upstate/Millipore
H3K4me3	16H10	Mouse (IgG)	10 µg purified	Kind gift from Hiroshi Kimura
H3K27me3	07-449	Rabbit (IgG)	50 µl	Upstate/Millipore
H3K36me3	13C9	Mouse (IgG, hybridoma)	20 µl	Kind gift from Hiroshi Kimura
H2A	07-146 (acidic patch)	Rabbit	-	Upstate/Millipore
H3	ab1791	Rabbit	-	Abcam Ltd, Cambridge, UK

Polycomb components				
Ring1B	A-20	Mouse (IgG, hybridoma)	50 µl	Kind gift from Haruhiko Koseki
Ezh2	pAb-039-050	Rabbit (IgG)	5 µl (5 µg)	Diagenode, Liège, Belgium
Controls				
Digoxigenin	Mouse (IgG)	10 µl (13 µg)	10 µl (13 µg)	Jackson ImmunoResearch Technologies, West Grove, PA
Anti-mouse IgM (µ chain)	Rabbit	10 µl	-	MP Biochemicals, Irvine, CA
Bridging antibodies				
Anti-Mouse IgM, µ Chain Specific	Rabbit	-	10 µg per 50 µl beads	Jackson ImmunoResearch Technologies
Anti-Mouse IgG+IgM (H+L)	Rabbit	-	10 µg per 50 µl beads	Jackson ImmunoResearch Technologies

Table 2.2 Antibodies used for Western analysis.

Antibody against	Clone	Raised in (isotype)	Working dilution	Origin
RNAPII				
RNAPII S2P	H5 (MMS-129R)	Mouse (IgM)	1:500	Covance, Princeton, NJ
RNAPII S5P	CTD4H8 (MMS-128P)	Mouse (IgG)	1:200,000	Covance
RNAPII S7P	4E12	Rat (IgG, hybridoma)	1:50	Kind gift from Dirk Eick
Non-phosphorylated S2 CTD residues	8WG16 (MMS-126R)	Mouse (IgG)	1:200	Covance
RPB1 N-terminus (amino acids 1-224)	H224 (sc-9001x)	Rabbit (IgG)	1:200	Santa Cruz Biotechnology, Santa Cruz, CA
Polycomb components				
Ring1B	A-20	Mouse (IgG, hybridoma)	1:500	Kind gift from Haruhiko Koseki
Ezh2	pAb-039-050	Rabbit (IgG)	1:1000	Diagenode, Liège, Belgium
Pluripotency markers				
Pou5f1	Sc-8628	Goat (IgG)	1:1000	Santa Cruz Biotechnology, Santa Cruz, CA
Sox2	Ab 158630	Rabbit (IgG)	1:2000	Abcam Ltd, Cambridge, UK
Nanog	Ab 80892	Rabbit (IgG)	1:300	Abcam Ltd, Cambridge, UK

Table 2.3 ChIP primers

Active genes	
b-actin promoter F	GCAGGCCTAGTAACCGAGACA
b-actin promoter R	AGTTTTGGCGATGGGTGCT
b-actin coding F	TCCTGGCCTCACTGTCCAC
b-actin coding R	GTCCGCCTAGAAGCACTTGC
Oct4 promoter F	GGCTCTCCAGAGGATGGCTGAG
Oct4 promoter R	TCGGATGCCCCATCGCA
Oct4 coding F	CCTGCAGAAGGAGCTAGAACA
Oct4 coding R	TGTGGAGAAGCAGCTCCTAAG
Sox2 promoter F	CCATCCACCCTTATGTATCCAAG
Sox2 promoter R	CGAAGGAAGTGGGTAAACAGCAC
Sox2 coding F	GGAGCAACGGCAGCTA
Sox2 coding R	GTAGCGGTGCATCGGT
Polr2a gene primers	
Rpb1 (-1kb) F	CCGTAAAGCTATTAGAGCACAGG
Rpb1 (-1kb) R	ATGCATAAGGCAGGCAAGAT
Rpb1 (-0.5kb) F	GTAACCTCTGCCGTTCAGGA
Rpb1 (-0.5kb) R	TTTCTCCCTTTCCGGAGATT
Rpb1 1 F	CAGGCTTTTTGTAGCGAGGT
Rpb1 1 R	GACTCAGGACTCCGAACTGC
Rpb1 2 F	TGGGTCAGTGATGCTGATGT
Rpb1 2 R	CTGGGGATCCACTTCCTGTA
Rpb1 3 F	CAGAGGGCTCTTTGAATTGG
Rpb1 3 R	GCATCAGATCCCCTTCATGT
Rpb1 4 F	CCAAGTTCAACCAAGCCATT
Rpb1 4 R	TCTTAACCGCTGAGCCATCT
Rpb1 5 F	TCCCAACTATACCCCGACAT
Rpb1 5 R	TGGTGAGCTTGGTGTGTAGG
Rpb1 6 F	TCTCCCACTTCTCCTGGCTA
Rpb1 6 R	CCGAGGTTGTCTGACCCTAA
Rpb1 (+0.6kb) F	TGCCCTTTTCTGGAGTGTCT
Rpb1 (+0.6kb) R	GCCAGGACTACACAGGCATT
Rpb1 (+2kb) F	GAGGGGCAGACACTACCAAA
Rpb1 (+2kb) R	AAAAGGCCAAAGGCAAAGAT
Rpb1 (+5kb) F	AATGCACAAACCCACACTCA
Rpb1 (+5kb) R	CGCTGAGTGCATTCTTGGTA
PRC-repressed genes	
Math1 promoter F	CCTTCTTTGACTGGGCAGAC
Math1 promoter R	ACTCGGAGATCGCACACC
Math1 coding F	CCAGTTGCCATTGCTTTAT
Math1 coding R	AGGATACTAGATTTGCAACATTCTT
Msx1 promoter F	ACAGAAAGAAATAGCACAGACCATAAGA

Msx1 promoter R	TTCTACCAAGTTCCAGAGGGACTTT
Msx1 coding F	AGATGGCCGCGAAAC
Msx1 coding R	CCAGAGGCACTGTAGAGTGA
HoxA7 promoter F	GAGAGGTGGGCAAAGAGTGG
HoxA7 promoter R	CCGACAACCTCATACCTATTCCTG
HoxA7 coding F	CTGGACCTTGATGCTTCTAACT
HoxA7 coding R	AGCCAGAGAAAGAGGGATTCTA
Gata4 promoter (-0.5 kb) * F	AAGAGCGCTTGCGTCTCTA
Gata4 promoter (-0.5 kb) * R	TTGCTAGCCTCAGATCTACGG
Gata4 coding (A) * F	TTGCACATTAACACCACACGTATA
Gata4 coding (A) * R	CCACCATTCAATTTTTAAGTCAAGTA
Silent genes	
Gata1 promoter F	AGAGGAGGGAGAAGGTGAGTG
Gata1 promoter R	AGCCACCTTAGTGGTATGACG
Gata1 coding F	TGGATTTTCCTGGTCTAGGG
Gata1 coding R	GTAGGCCTCAGCTTCTCTGTAGTA
Myf5 promoter F	GGAGATCCGTGCGTTAAGAATCC
Myf5 promoter R	CGGTAGCAAGACATTAAAGTTCCGTA
Myf5 coding F	GATTGCTTGTCCAGCATTGT
Myf5 coding R	AGTGATCATCGGGAGAGAGTT

Sequences are reported in 5' to 3' orientation. (*) Indicates primers with two designations for their use, in promoter and coding region analyses and in the detailed mapping of single genes.

Table 2.4 Expression primers used to detect spliced transcripts.

House-keeping genes	
b-actin F	TCTTTGCAGCTCCTTCGTTG
b-actin R	ACGATGGAGGGGAATACAGC
UBC F	AGGAGGCTGATGAAGGAGCTTGA
UBC R	TGGTTTGAATGGATACTCTGCTGGA
G6PD (5)	CGACAGTTGATTGGAGCTCTG
G6PD (3)	AGCCACATGAATGCCCTGCAC
Pluripotency genes	
Oct4 F	ACCTCAGGTTGGACTGGGCCTA
Oct4 R	GCCTCGAAGCGACAGATGGT
Nanog F	AATTCTGGGAACGCCTCAT
Nanog R	TTGTTTGGGACTGGTAGAAGAATC
Sox2 F	CATGTGAGGGCTGGACTGCG
Sox2 R	GCTGTCGTTTCGCTGCGG
PRC-repressed genes	
Math1 F	GGAGAAGCTTCGTTGCACGC
Math1 R	GGGACATCGCACTGCAATGG
Nkx2.2 F	TGTGCAGAGCCTGCCCCTTAA

Nkx2.2 R	GCCCTGGGTCTCCTTGTCAT
Msx1 F	GCCTCTCGGCCATTTCTCAG
Msx1 R	CGGTTGGTCTTGTGCTTGCG
Nkx2.9 F	GGCCACCTCTGGACGCCTCG
Nkx2.9 R	GCCAGCTGCGACGAGTCTGC
Mash1 F	TGGAGACGCTGCGCTCGGC
Mash1 R	CGTTGCTTCAATGGAGGCAAATG
Cdx2 F	CAGCCGCCGCCACAACCTTCCC
Cdx2 R	TGGCTCAGCCTGGGATTGCT
HoxA7 F	AAGCCAGTTTCCGCATCTACC
HoxA7 R	GTAGCGGTTGAAATGGAATTCC
Flk1 F	AGGGGAACTGAAGACAGGCTA
Flk1 R	GATGCTCCAAGGTCAGGAAGT
Gata4 F	GAGGCTCAGCCGCAGTTGCAG
Gata4 R	CGGCTAAAGAAGCCTAGTCCTTGCTT
HoxB1 F	AGGAATCGCCTTGCTCG
HoxB1 R	GTGAAGTTTGTGCGGAGACC
HoxD1 F	GCCCACAGCACTTTCGA
HoxD1 R	CTGAAATTTGTGCGGATGG
HoxD13 F	GTGTACTGTGCCAAGGATCA
HoxD13 R	TGTCCGGCTGGTTTAAAG
Silent genes	
Gata1 F	GTCCTCACCATCAGATTCCACAG
Gata1 R	AGTGGATACACCTGAAAGACTGGG
Myf5 F	GGAGATCCTCAGGAATGCCATCCGC
Myf5 R	GACGTGATCCGATCCACAATGCTGG

Sequences are reported in 5’ to 3’ orientation.

3. Proteome ChIP (pChIP) as a tool to dissect chromatin-bound proteome (Optimizing pChIP method)

3.1. Research motivation

Chromatin proteins provide a scaffold for DNA packaging and basis for epigenetic regulation and gene expression. With the advent of chromatin proteomics and methods, the field is increasingly identifying large cohorts of proteins and capturing novel biological interactions that would not be possible with conventional methods.

My aim in this project was to investigate the composition of proteome of RNAPII-bound chromatin bound proteome and further dissect the proteome dependencies to specific RNAPII modifications. Therefore first I aimed to understand diversity of proteins obtained from different chromatin preparations using mass spectrometry in mES cells. My aim was to develop an unbiased method called 'Proteome-ChIP' that parallels DNA-ChIP and captures the cohorts of interactions occurring on chromatin along with protein of interest (i.e. RNAPII) in mES cells. Moreover to perform pChIP on distinct RNAPII modifications (RNAPII-S5p and RNAPII-S7p) to capture their chromatin bound proteome thereby identifying novel proteins and also uncovering known associations.

All of the MS experiments were carried out in collaboration with Dr. Bram Snijders at Proteomics facility at MRC-CSC, who helped with sample pre-processing for MS and performed all MS run time operations and quantification of MS spectra. Dr. Hiroshi Kimura (Osaka University, Japan) shared initial conditions and advice for protein extraction after ChIP.

3.2. Stem cells and regulation of gene expression

Embryonic stem (ES) cells are derived from the inner cell mass (ICM) cells of the blastocyst stage of early embryo. ES cells retain the ability to differentiate

into somatic lineages of all three germs layer, a characteristic referred to as pluripotency. ES cells can also grow indefinitely in culture under appropriate conditions (self renewing), thereby making them invaluable model system to study early development and potential for therapeutic possibilities (Jaenisch and Young 2008). The gene expression in ES cells is quite dynamic and well regulated by specific Transcription factors (TFs), chromatin modifiers and regulatory modules that interact on chromatin and cascade a range of downstream processes essential not only for pluripotency and self-renewal but also for normal function of cellular processes.

The ES cell chromatin is tightly and dynamically regulated owing to architectural proteins that contribute to genome plasticity (Meshorer *et al.* 2006; Melcer and Meshorer 2010). The ES cell chromatin provides the platform for TFs and master regulators (e.g. Oct4, Sox2, Nanog etc.) to interact and direct chromatin states whereby cohorts of interactions regulated gene expression. The ES cell chromatin is thought to be transcriptionally hyperactive and central to this RNA polymerase II (RNAPII) the proteins that transcribes through on chromatin leading to production of stable and mature mRNA as well as many structural and non-coding RNAs(Efroni *et al.* 2008; Young 2011).

3.3. RNAPII transcription in mES cells.

In mES cells, we have shown that RNAPII not only binds and transcribes active genes defined by open chromatin architecture, but is also known to associate with PRC-repressed genes (Polycomb repressed genes) characterised by bivalent chromatin modifications (both open and closed chromatin marks) and unusual RNAPII variant (Stock *et al.* 2007b; Brookes *et al.* 2012). At active genes, RNAPII undergoes dynamic and sequential phosphorylation on its carboxy-terminal domain (CTD) at serine's on position 5,7 and 2 respectively which correlate with transcription initiation, transition to elongation and transcription elongation respectively. The successful round of

transcription additionally involves de-phosphorylation, recycling and unloading of RNAPII from chromatin for next round of transcription cycle (Brookes and Pombo 2009b).

Remarkably at PRC-repressed genes, RNAPII initiates transcription i.e. S5p and transcribes across coding regions in the absence of S7p or S2p but without expression or production of stable mRNA. The PRC-repressed genes constitute almost 25% of the Refseq genes and encode important developmental regulator genes that are activated upon early steps of lineage specification (Brookes *et al.* 2012). Therefore the need to explore the chromatin bound RNAPII-proteome in mES cells became apparent as a means of identify additional activities associated with RNAPII-S5p at PRC-target genes.

3.4. Results

3.4.1. Proteome complexity is reflected by different chromatin preparations.

To understand the complexity of proteins that can be extracted from different chromatin preparations, I started by preparing three different types of chromatin and comparing their pattern of DNA enrichment and proteome composition. *Native chromatin* is prepared by micrococcal nuclease digestion of cell nuclei and captures resolution at level of mono-, di- and tri-nucleosomes (O'Neill and Turner 2003). It is primarily used for analysis of histones and their post translationally modified isoforms. *Fixed chromatin* (or Fixed nuclear preparation) is prepared by formaldehyde crosslinking of cells, prior to nuclear isolation, to preserve DNA-protein and protein-protein interactions (within 2Å spacer arm) and is mainly used for capture transcription factor (TF) binding across the genome. Fixed chromatin is essentially a formaldehyde-crosslinked total nuclear extract that preserves DNA-protein, protein-protein and RNA-protein interactions (within the spacer arm distance). *Double-crosslinked chromatin* is an adaptation of fixed

chromatin wherein two crosslinking agents (formaldehyde and EGS) are used in tandem to capture direct and indirect DNA-associated protein interactions (Zeng *et al.* 2006).

After DNA extraction from the three chromatin preparations and electrophoresis on an Agarose gel (Fig. 3.1), I observed that native chromatin mostly consisted of mono-nucleosome fragments and to some extent di- and tri-nucleosomes (fragment size range 145 - 300 bp). Fixed and double-crosslinked chromatin produced DNA fragments of typical size range 0.25-1.2kb.

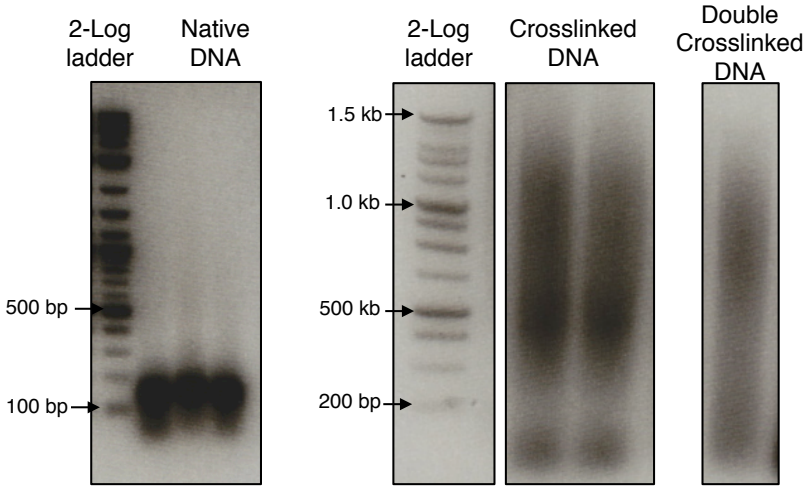


Figure 3.1 Diversity of DNA fragments in different chromatin preparations separated by Agarose gel electrophoresis. DNA from native chromatin, fixed chromatin and double-crosslinked chromatin was extracted and separated by electrophoresis on 1.2% Agarose gel. DNA extraction was performed as described (section 2.2.9). The DNA fragment range from native chromatin reflects the size distribution of mono- and di-nucleosomes, while the fixed and double-crosslinked chromatin show a range of 300bp-1.2kb.

Before investigating the proteome composition in different chromatin preparations, I first optimised conditions for reverse-crosslinking of proteins from fixed chromatin thereby facilitating the reduction and denaturation of proteins. Unlike DNA detection by PCR, proteins cannot be amplified, so the efficiency of extraction and the protein yield are crucial for successful protein identification and quantification. I optimised reverse crosslinking (RCL)

conditions by treating fixed chromatin samples under a range of conditions with varying temperature and time (Fig. 3.2).

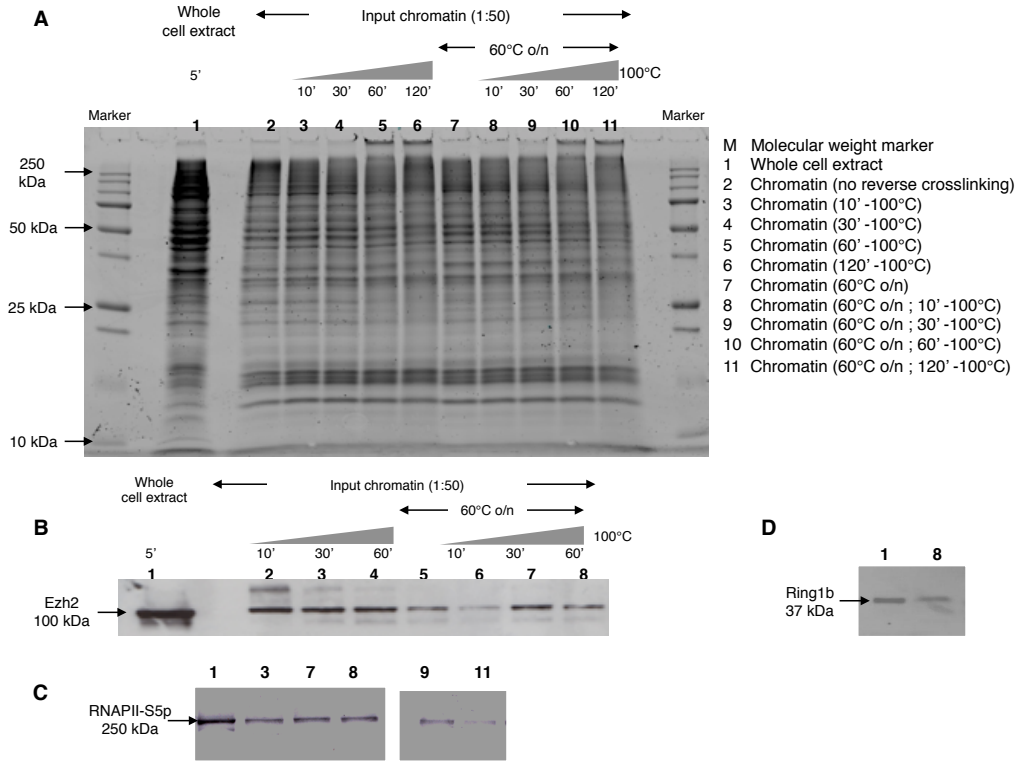


Figure 3.2 Reverse crosslinking conditions to extract proteins from fixed chromatin. Fixed chromatin was re-suspended in custom Laemmli buffer and subjected to range of conditions (with varying temperature and time) to effectively reverse crosslink and elute proteins. (A) Coomassie stained gel with whole cell extract proteins and chromatin proteins under different RCL conditions. Minor effects are noticed at the level of total protein analyses, although it is noticeable that longer incubations at 100°C give less definition in the protein banding pattern (B, C and D) Western blotting of proteins (Ezh2, RNAPII-S5p and Ring1b) from whole-cell-extract and fixed chromatin preparations reverse crosslinked under a range of conditions. Different proteins have different sensitivity to the RCL condition. I chose condition 8 (60°C o/n and 100°C for 10min) for further experiments.

After Coomassie staining of chromatin proteins, I observed that most of the RCL conditions gave a similar pattern of distribution of proteins. However, proteins seemed to degrade when exposed to longer incubation and at higher temperature (Fig. 3.2A, lanes 5, 6, 10 and 11). Probing for specific proteins by western blotting, it was apparent that the different RCL conditions had slightly different effects on proteins. For example, Ezh2 protein was best extracted in lanes 4 and 7 (Fig. 3.2B; lanes 2 and 3 show a doublet not present in whole

cell extract, consistent with insufficient RCL). RNAPII-S5p was extracted optimally in lanes 7 and 8 (Fig. 3.2C). We decided to fix our RCL conditions (60°C o/n and 100°C for 10min; lane 8) keeping in mind that specific proteins might have a different optimal RCL condition.

Next, I compared the proteome composition of the three different chromatin preparations (native, fixed and double-crosslinked) by Coomassie staining of chromatin extracts after separation by SDS-PAGE (Fig. 3.3). As expected, I observed that native chromatin is enriched particularly for histones, but interestingly a number of proteins are also detected in the larger size range of 25 to 250 kDa). As anticipated, fixed chromatin is highly enriched for a range of proteins of different molecular weights, including core histones. Detection of proteins on the stacking portion of the gel suggests incomplete reverse-crosslinking of proteins. Double crosslinked chromatin gave rise to the most intense Coomassie staining, representing the most diverse range of proteins with most detectable enrichment. As with fixed chromatin, stacking was observed with double-crosslinked chromatin.

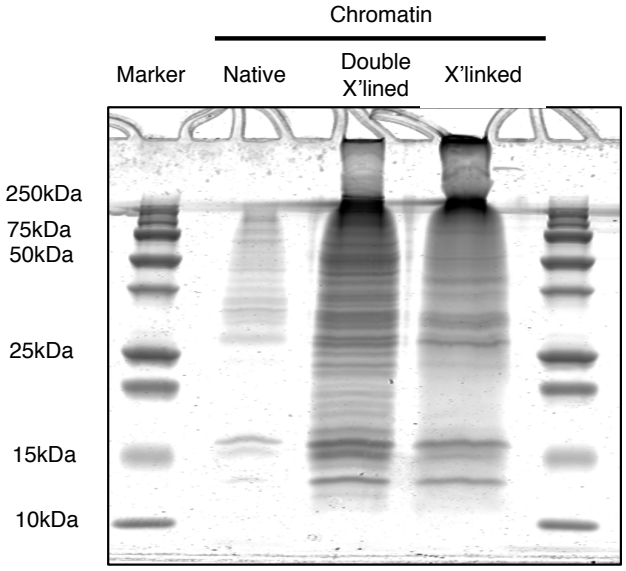


Figure 3.3 Diversity of proteins present in different chromatin preparations visualised by Coomassie staining. Native proteins were reduced and denatured by adding custom prepared Laemmli buffer (Section 2.4.1, 1:1 ratio by volume). Fixed and double-crosslinked chromatins were first reverse crosslinked (60°C o/n and 100°C 10min) and eluted in custom Laemmli buffer (1:1 by volume).

To explore the specific enrichment or depletion of proteins between native and fixed chromatin and their function, I analysed the proteins from native and fixed chromatin by mass spectrometry (MS) and performed Gene Ontology (GO) analyses to ask which protein groups were shared and/or selectively enriched and their biological function (Fig. 3.4). MS analyses robustly identified 1320 and 2455 proteins in native and fixed chromatin, respectively, with ~1000 proteins overlapping between the two datasets. Nucleosomes (histones) and chromatin components were enriched and shared in both datasets. I performed GO analysis to identify biological processes shared between the datasets and also enriched in one preparation over another. Shared proteins were enriched for GO terms 'nucleosome assembly', 'RNA splicing', 'mRNA processing' and 'translation' (p-values $<2x10^{-6}$). We expected histones to be shared between the two chromatin methods, however enrichment of mRNA processing and RNA splicing terms along with histone terms suggest that we may also enrich for chromatin proteins involved in active transcription. Interestingly, proteins enriched only in native conditions (344 proteins) are involved in processes including G-protein signalling, blastocyst formation and positive regulation of transcription and include proteins fundamental for stem cell pluripotency and self renewal (Oct4, Sox2, Dppa2 and Dppa4). Fixed chromatin represents essentially a total nuclear extract, and it is perhaps not surprising that a larger number of proteins is identified enriched only in fixed chromatin (1322 proteins) which are involved in diverse biological processes including signalling, transport and proteolysis.

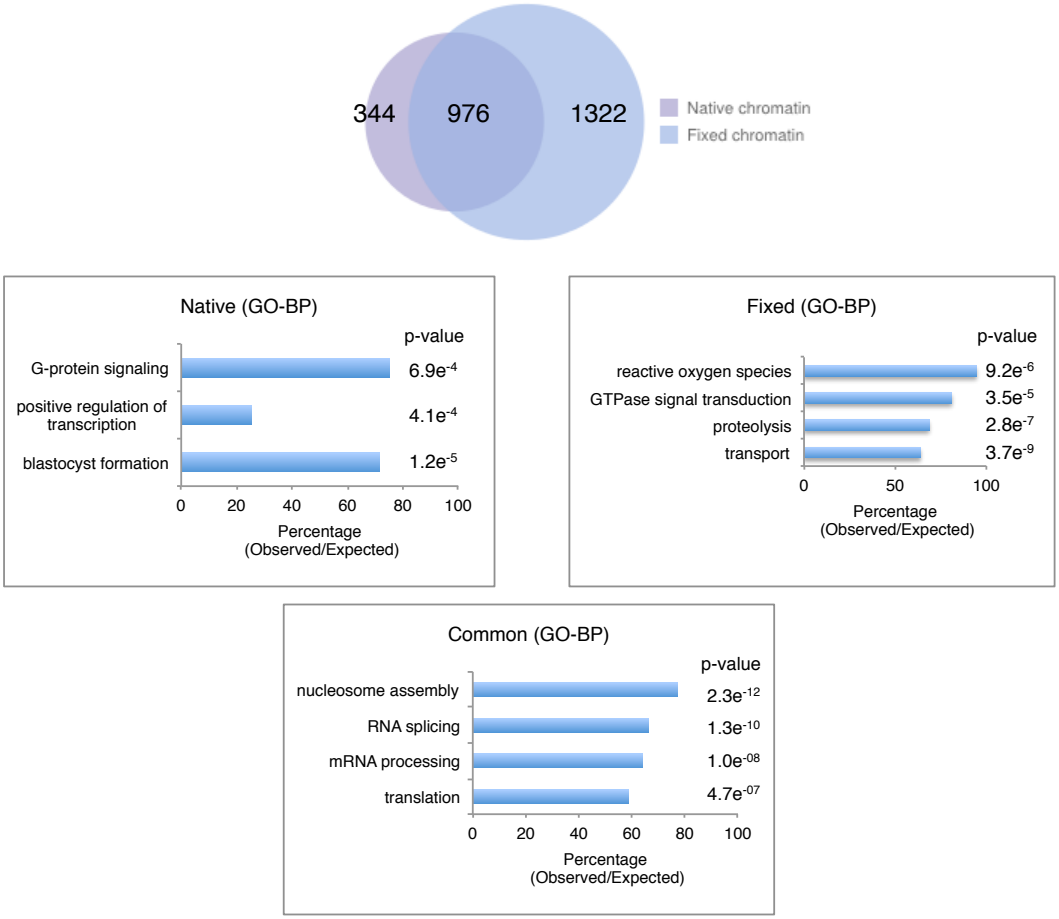


Figure 3.4 Comparing the proteome composition of native chromatin and fixed chromatin. MS identified 1320 proteins from native chromatin (signal intensity > 20,000) and 2455 proteins from fixed chromatin (mascot score > 50). Significant overlap (976 proteins) was observed between the datasets. Gene ontology (GO) for biological processes (BP) was performed for proteins common in both preparation and proteins identified only specific chromatin preparations (p-value < 10-4).

In summary, the comparisons between native, fixed and crosslinked chromatin show the expected increase in protein complexity (Fig. 3.4) whereas comparisons of different RCL conditions demonstrate that it is possible to successfully extract proteins for MS analyse from crosslinked chromatin (Fig. 3.2). Formaldehyde crosslinked (fixed) chromatin was chosen as a starting material for further analyses by proteome-ChIP (pChIP), to identify the cohorts of proteins that co-associate with RNAPII bound to chromatin. First, our laboratory has developed a DNA-ChIP protocol optimised

to detect the chromatin occupancy of RNAPII modifications using fixed chromatin. Second, many proteins are lost in native chromatin preparations. Third, formaldehyde crosslinks can capture robust and transient interactions within 2Å spacer arm. Fourth, we could demonstrate the feasibility of detecting proteins by MS from formaldehyde-crosslinked chromatin, after optimising the RCL conditions.

To check the specificity and robustness of our detection of chromatin proteins from crosslinked chromatin, I decided to compare our MS dataset from fixed chromatin from mES-OS25 cells, with published datasets for fractionated nucleoplasmic, cytoplasmic and chromatin proteins from a different mES cell line (Graumann *et al.* 2008) (Fig. 3.5). We observe a 60% overlap (1331 proteins out of 2191 proteins) with published chromatin fraction proteins. In addition, our fixed chromatin dataset contains very few cytoplasmic (158 proteins) and nucleoplasmic proteins (198 proteins). The high specificity and depth of our chromatin MS dataset is highlighted by the fact that 22% (496 proteins) of our chromatin proteins were not detected by the previous study. Moreover, we enrich for important chromatin remodellers (Nanog, Bap18, Carm1, Setd3, CCR4-NOT complex), metabolic proteins (Arglu1, Prmt3, Hk3) and transcription associated proteins (mediator subunits, integrator subunits, RNAPII phosphatases, RNA binding proteins and splicing factors).

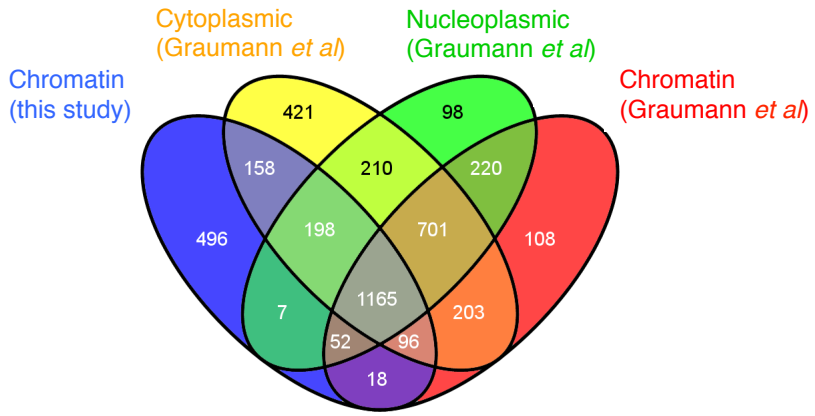


Figure 3.5 Specificity of our fixed chromatin in comparison with different cellular fractions. MS robustly identified 2191 proteins from our fixed chromatin from mES-OS25 cells. Proteins from published MS analysis of mES cells (Graumann *et al.* 2008) identified 3152, 2651 and 2564 proteins from native cytoplasmic, nucleoplasmic and chromatin fractions respectively.

Taken together, these analyses of total chromatin suggest the feasibility of analysing the proteome of fixed chromatin. In the next sections, I combine the protein extraction steps optimised here with protein extraction from chromatin immunoprecipitated with different RNAPII antibodies using an optimised ChIP protocol (Stock *et al.* 2007b).

3.4.2. DNA-ChIP using RNAPII antibodies.

Before optimising conditions for identifying the proteome of chromatin by pChIP using RNAPII antibodies, I started by reproducing previous DNA-ChIP analysis of RNAPII occupancy across a panel of genes in mES cells (Stock *et al.* 2007b; Brookes *et al.* 2012). I used three highly specific and well-characterised antibodies directed against phosphorylation on S5, S2 and S7 residues on the RNAPII CTD, which reflect distinct stages of the transcription cycle. RNAPII-S5p marks transcriptional initiation and is found highly enriched at promoter of actively transcribed genes and also at Polycomb repressed genes. RNAPII-S7p transitions RNAPII from transcription initiation to transcription elongation and is enriched at promoters of actively transcribed genes. RNAPII-S2p occurs downstream of initiation, marks transcriptional

elongation and is associated with coding regions of actively transcribed genes.

Consistent with the published data (Stock *et al.* 2007b; Brookes *et al.* 2012) in my DNA-ChIP results (Fig. 3.6), RNAPII-S5p was found enriched at promoters and coding (+2kb downstream of TSS) regions of active and PRC-repressed genes. RNAPII-S7p was enriched at promoters and coding regions of active genes consistent with role in active transcription. RNAPII-S2p is enriched only at the active genes, with enrichment at coding regions much higher than at promoters. Mock ChIP (control immunoprecipitation using an antibody against plan steroid, Digoxigenin) demonstrates the specificity of our ChIP protocol, with no detectable DNA enrichment compared to ChIP with RNAPII-S5p, -S7p or -S2p antibodies.

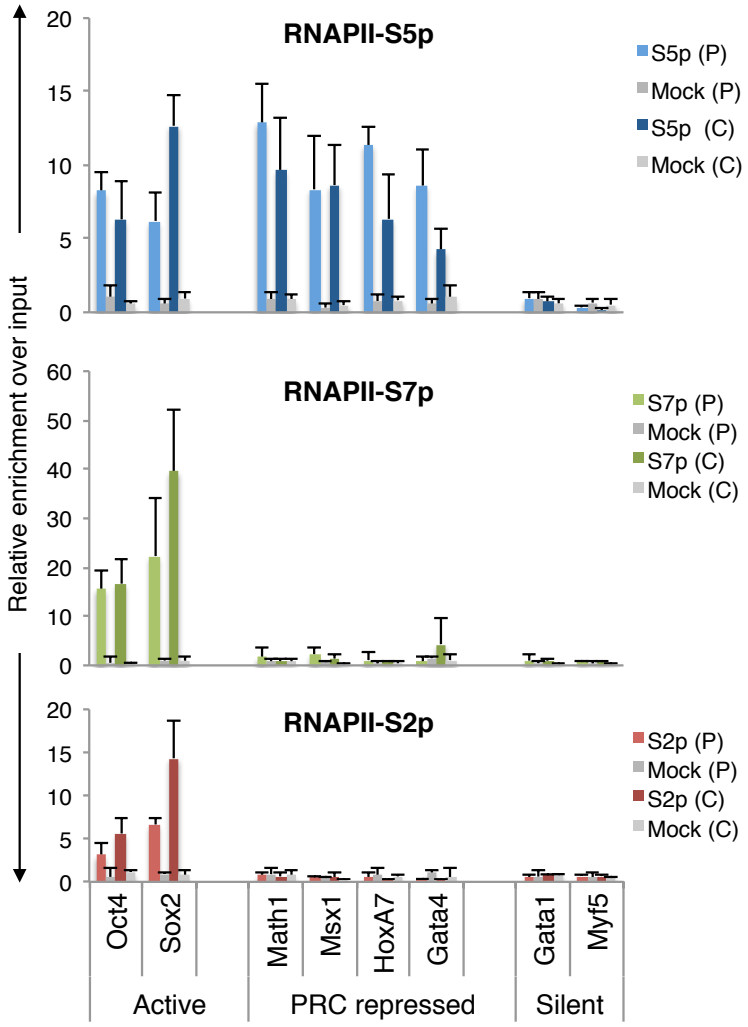


Figure 3.6 Occupancy of different RNAPII modifications across a panel of active, PRC-repressed and silent genes in mES cells. Occupancy of RNAPII-S5p (light and dark blue), -S7p (light and dark green) and -S2p (light and dark red) as measured by DNA-ChIP and qRT-PCR at promoters (P, light) and coding (C, dark) regions of panel of two active, four PRC-repressed and two inactive genes. Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. Mean and standard deviation are representative of 3 independent biological replicates (2 replicates for S2p).

To further demonstrate the quality of the ChIP experiments, I also measured the chromatin occupancy of S5p, S7p and S2p across a single gene locus, the polr2a gene, spanning ~28kb. I observed RNAPII-S5p highly enriched at promoter and present throughout coding regions (Fig. 3.7). RNAPII-S7p is

highly enriched at promoters and RNAPII-S2p enrichment occurs predominantly downstream of promoters and towards end of gene.

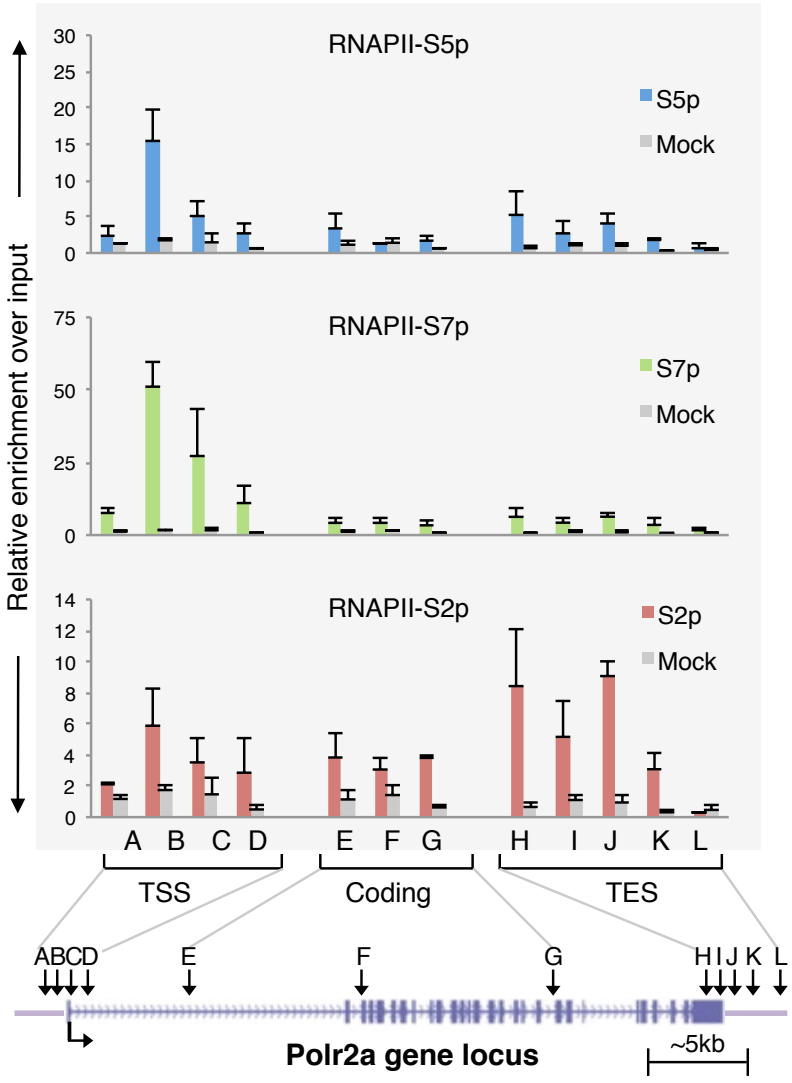


Figure 3.7 Occupancy of RNAPII modifications across a single gene locus in mES cells. Occupancy of RNAPII-S5p (blue), -S7p (green) and -S2p (dark red) as measured by DNA-ChIP and qRT-PCR across a single gene locus (Polr2a – 28kb) spanned by 12 spatially positioned primers (primers kindly designed by I. de Castro; our lab). Grey bars represent background enrichment levels as measured by control immunoprecipitation (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. Mean and standard deviation are representative of 3 independent biological replicates.

These analyses show that the DNA-ChIP detects specific RNAPII modifications at the appropriate locations, and with good and reproducible yields comparable with previous analyses in the laboratory (Stock *et al.*

2007b). In the next section, I combine the ChIP procedure with the RCL extraction of proteins and MS analyses, towards identifying the proteome associated with RNAPII complexes bound to chromatin in ES cells.

3.4.3. Proteome-ChIP (pChIP): Overview and optimization

With an aim to identify the RNAPII-bound chromatin proteome, I first developed and optimised the pChIP protocol to effectively enrich and extract proteins that bind to RNAPII on chromatin. With pChIP, we aim to capture the protein complexes that are associated with chromatin at the same time as RNAPII, *i.e.* all cohorts of interactions irrespectively of whether they are direct or indirect associations with RNAPII, and whether they are mediated by DNA-protein, protein-protein or RNA-protein interactions. Briefly, pChIP is similar to DNA-ChIP, wherein mES cells are *in-vivo* crosslinked with formaldehyde to fix chromatin interactions, before nuclei are purified and sonicated to produce fixed chromatin. Using highly specific RNAPII antibodies, we immunoprecipitate chromatin, reverse crosslink chromatin and elute proteins to be further analysed either by western blotting or by high-throughput MS approaches (Fig. 3.8).

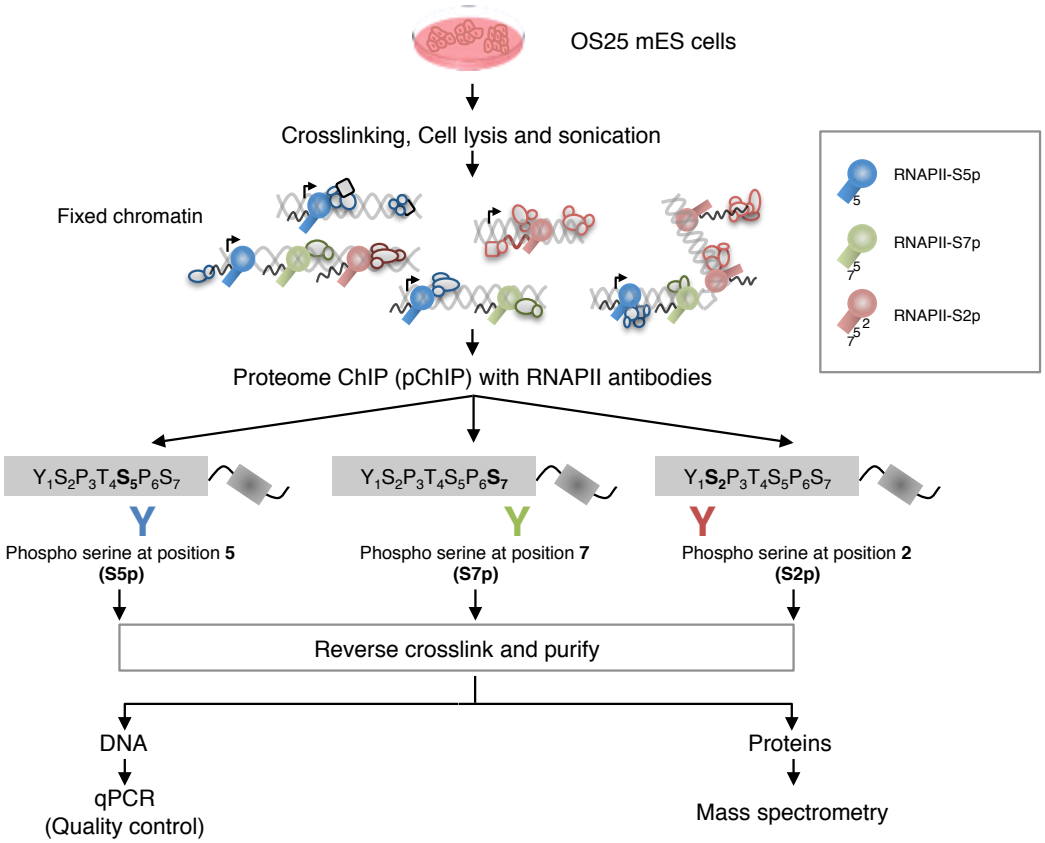


Figure 3.8 Overview of steps involved in Proteome-ChIP (pChIP). mES cells are crosslinked, cell lysis is followed by nuclei purification and sonication to produce fixed chromatin. Using highly specific antibodies, immunoprecipitation is performed and DNA and proteins are reverse crosslinked in parallel. Occupancy of RNAPII modifications is measured by qRT-PCR. Extracted proteins can be identified by MS or individually probed by western blotting.

To test the conditions for pChIP, I started by performing RNAPII-S5p ChIP (along with mock ChIP), reverse crosslinking and analysing the protein composition using Coomassie staining of the SDS-PAGE gel (Fig. 3.9). There was clear Coomassie staining in the RNAPII-S5p pChIP (lane 4) in contrast with the Mock pChIP (lane 5) where little or no staining was detected. Proteins bands were not clearly visible in the RNAPII-S5p ChIP lane, either due to presence of DNA in protein sample or incomplete RCL that would prevent the definition of specific protein bands and would favour stacking of chromatin at the top of the lane. Western blotting for RNAPII-S5p further confirmed the enriched of RNAPII in the respective pChIP sample and not in mock pChIP.

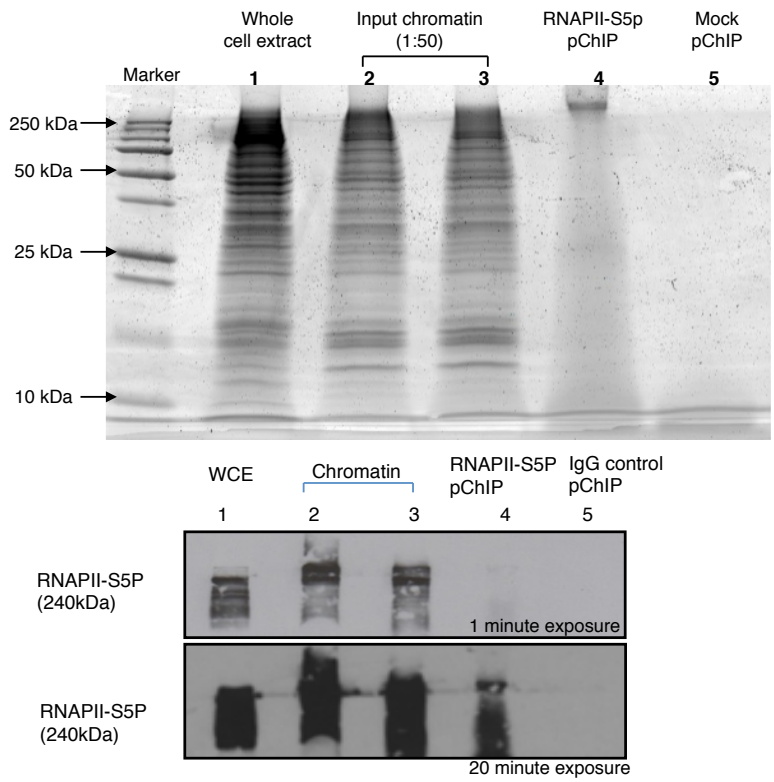


Figure 3.9 Western blot analysis of RNAPII-S5p pChIP. pChIP was performed for RNAPII-S5p with mock in the absence of primary antibody, before samples separated by SDC-PAGE and stained with Coomassie to visualize distribution of proteins. Staining was observed in RNAPII-S5p but protein bands were not defined in comparison with WCE (whole cell extract) or input chromatin. Mock pChIP was clean and with little or no Coomassie staining. Western blotting for RNAPII-S5p further confirmed specificity of pChIP as RNAPII-S5p was observed in all lanes except mock pChIP.

To improve the visualisation of proteins after SDS-PAGE and the protein extraction, I tested the effect of using different DNase treatments before elution of chromatin from the immunoprecipitating beads, sonication of the protein extract in Laemmli buffer and protein precipitation (Fig. 3.10). With most of these conditions, visualization was only partially improved, and at the cost of reduced protein amounts or increased processing steps. We chose to include a DNase (Benzonase) treatment at the last stage of immunoprecipitation, as standard condition to digest DNA before elution of proteins from the beads.

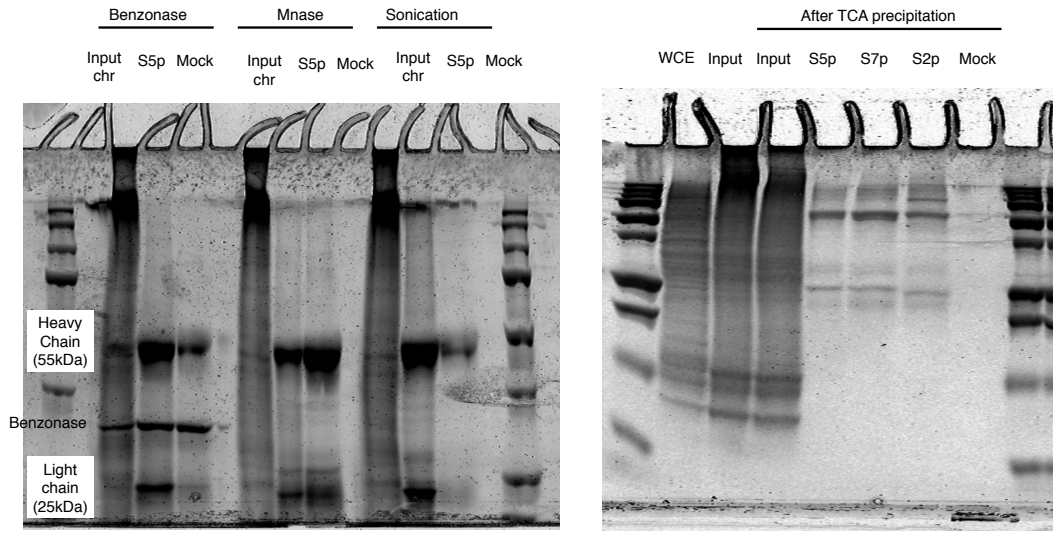


Figure 3.10 Testing conditions to improve protein visualisation after Coomassie staining. DNase (Benzonase, 25 units/final volume; MNase, 2 units/final volume) and sonication (10min 30sec on-off) were used to further fragment DNA in input, RNAPII-S5p and mock pChIP resulting in slightly improved visualization. TCA precipitation improved visualization of pChIP samples (S5p, S7p, S2p and Mock) but with reduced protein recovery.

As a summary, the sample processing steps chosen for subsequent pChIP experiments up to MS analysis are briefly described in Fig. 3.11.

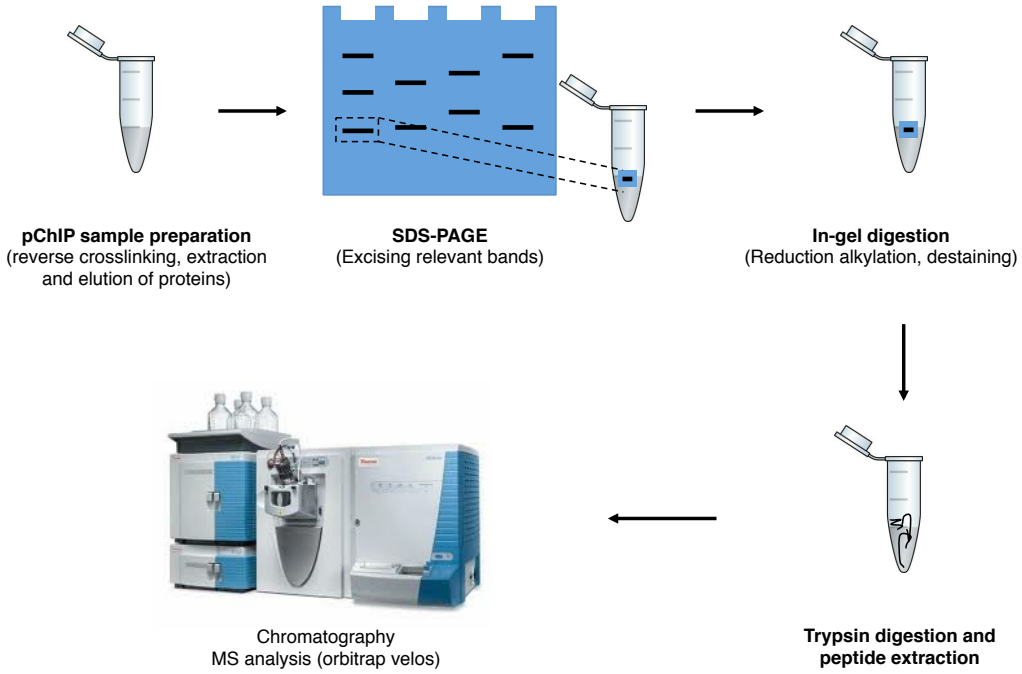


Figure 3.11 Samples processing steps for pChIP before MS data acquisition. Briefly samples are run on SDS-PAGE gel and Coomassie stained to visualise protein distribution. Accordingly complete gel or appropriate sections are cut and processed for in-gel reduction, alkylation and destaining. Proteins are digested by trypsin and tryptic peptides are extracted for chromatographic separation. Finally the different fractions are injected and analysed by tandem MS/MS. The spectra for peptides are generated and database search facilitates identification of peptides and therefore proteins present in sample.

3.4.4. RNAPII-S5p pChIP

To identify the cohorts of proteins interacting with RNAPII-S5p, I performed pChIP with RNAPII-S5p antibodies (along with mock pChIP) and analysed these two samples together with input fixed chromatin by MS. RNAPII-S5p is found predominantly at promoters but also at coding regions of both active and PRC-repressed genes (Fig. 3.5). MS analyses were performed at the CSC MS facility. After removal of known MS contaminants, 730 proteins were identified in RNAPII-S5p pChIP, 75 proteins were identified in mock pChIP and 2490 proteins were identified in input chromatin (Fig. 3.12).

To measure the specificity of pChIP, I compared proteins identified in RNAPII-S5p and mock pChIP and represented the results in a Venn diagram. Reassuringly, 96% of pChIP RNAPII-S5p proteins (706) were not detected in mock pChIP. I next asked whether RNAPII-S5p pChIP specifically enriched for proteins involved in RNAPII transcription and co-associated processes by comparing the MS datasets obtained with input fixed chromatin and RNAPII-S5p pChIP. Using the MS with intermediate depth, of 2490 proteins detected in input chromatin, RNAPII-S5p pChIP specifically enriched for 545 proteins (overlap – 75% of RNAPII-S5p proteome). Performing gene ontology (GO) analysis, I observed that the proteome enriched for with RNAPII-S5p pChIP was enriched for ‘mRNA’, ‘RNA processing” terms and ‘chromatin remodellers’ (Fig. 3.12), while the input chromatin was composed of various nuclear processes. The identification of few proteins in RNAPII-S5p only (185 proteins) reflects on incomplete protein sequencing depth of the more

complex input chromatin. A few examples of proteins enriched in RNAPII-S5p are listed in the Fig. 3.12.

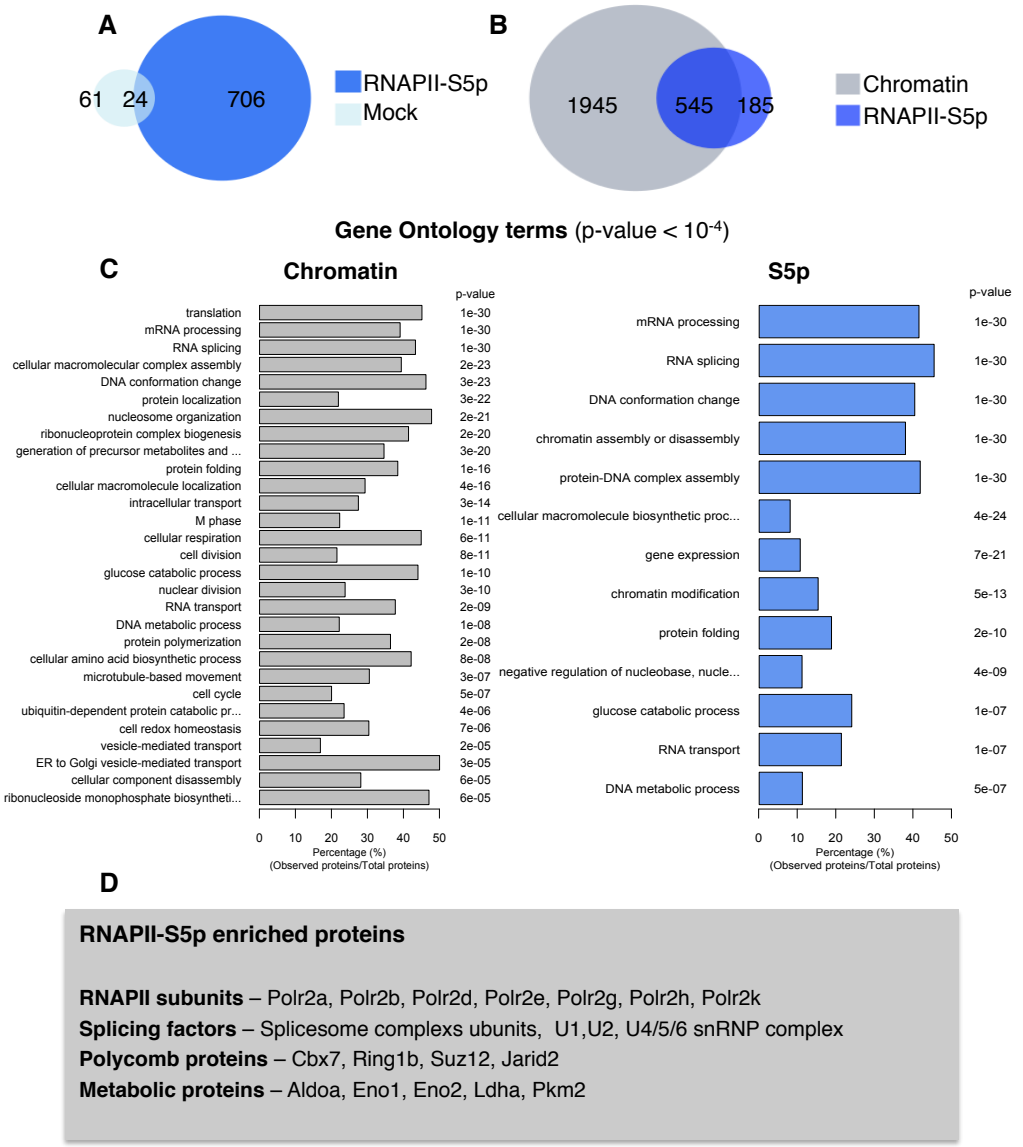


Figure 3.12 Specificity and robustness of proteins detected by RNAPII-S5p pChIP. (A) Specificity of RNAPII-S5p pChIP (dark blue) was measured by comparing with mock pChIP (light blue) in a Venn. (B) Robustness and specific enrichment of RNAPII-S5p over input chromatin proteins (grey) is displayed as Venn diagram. (C) Significant GO terms enriched in input chromatin and RNAPII-S5p pChIP. (D) Examples of proteins enriched in RNAPII-S5p.

3.4.5. RNAPII-S7p pChIP

To further explore the specificity of the chromatin proteome associated with RNAPII, I next performed pChIP with RNAPII-S7p (along with mock pChIP). RNAPII-S7p marks transition from transcription initiation to elongation and is found enriched at promoters of actively transcribing genes (Fig. 3.5). After RNAPII-S7p pChIP, 357 proteins were identified and only 26 proteins were also detected in mock pChIP highlighting the specificity of RNAPII-S7p pChIP (Fig. 3.13). Comparing RNAPII-S7p with input fixed chromatin, 67% of proteins were specifically enriched from input and GO analysis confirmed enrichment of mRNA and RNA processing terms (Fig. 3.13). RNAPII-S7p only proteins (110 proteins) are attributed to low protein sequencing depth of input chromatin as with RNAPII-S5p MS run.

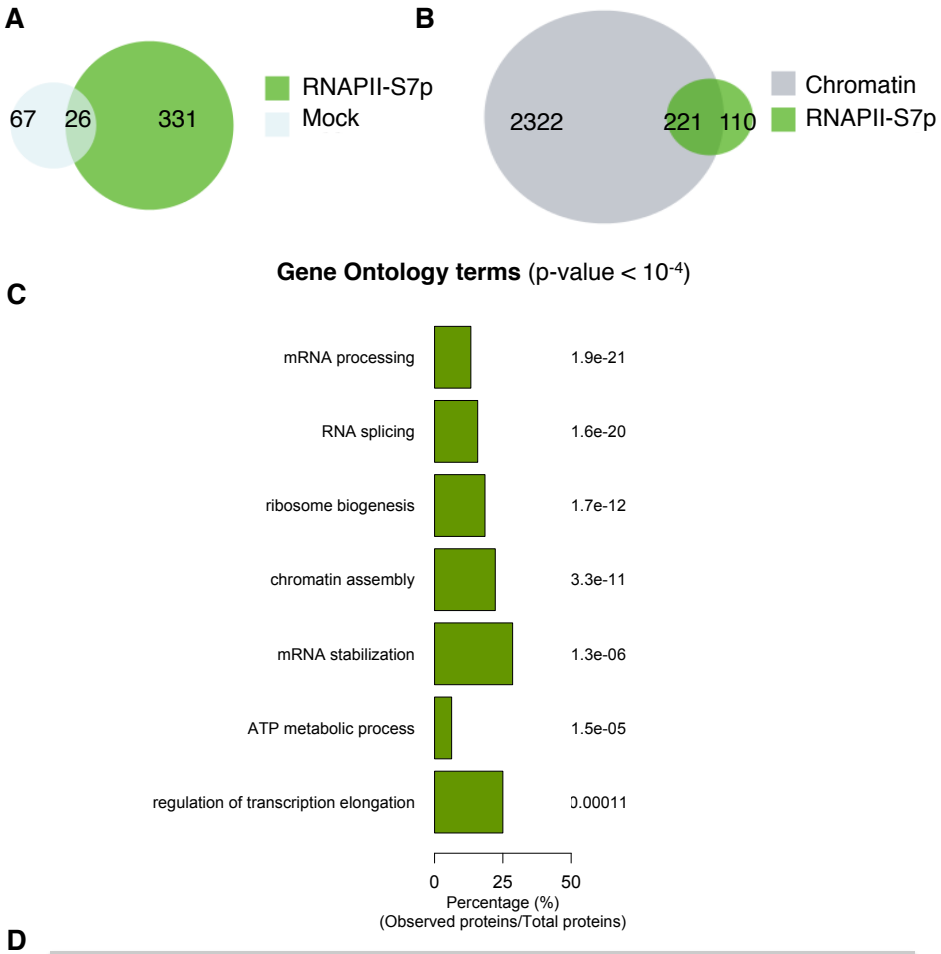


Figure 3.13 RNAPII-S7p pChIP. (A) RNAPII-S7p (green) enriches for proteins not identified in mock pChIP (grey) as plotted in a Venn diagram. (B) RNAPII-S7p (green) robustly and specifically enriches for proteins over input chromatin (grey). (C) Significant GO terms enriched in RNAPII-S7p pChIP. (D) Examples of proteins enriched in RNAPII-S7p.

3.4.6. pChIP between different RNAPII modifications highlights the proteome composition (RNAPII-S5p versus RNAPII-S7p)

To probe for the specificity of the proteome identified by pChIP with different RNAPII antibodies, I compared the two MS datasets obtained for pChIP with antibodies against RNAPII-S5p and S7p. In this label-free MS analysis, only

qualitative information on identified proteins is assessed i.e. a protein identified can be enriched/depleted between two samples whereas no evaluation can be done for proteins not identified. With this background, I asked how many proteins were identified in both RNAPII-S5p and RNAPII-S7p pChIP.

Interestingly 236 proteins were found common in the datasets, which included RNAPII subunits, histones, components of transcriptional co-processing machinery and splicing factors (Fig. 3.14). Unique proteins only identified in the RNAPII-S5p pChIP were 363 proteins, and in the RNAPII-S7p pChIP 140 proteins. Reassuringly Polycomb proteins are consistently detected only with RNAPII-S5p and not with S7p, whereas Transcription associated factors (TAFs) and histone proteins are detected in the S7p, but not S5p pChIP MS datasets.

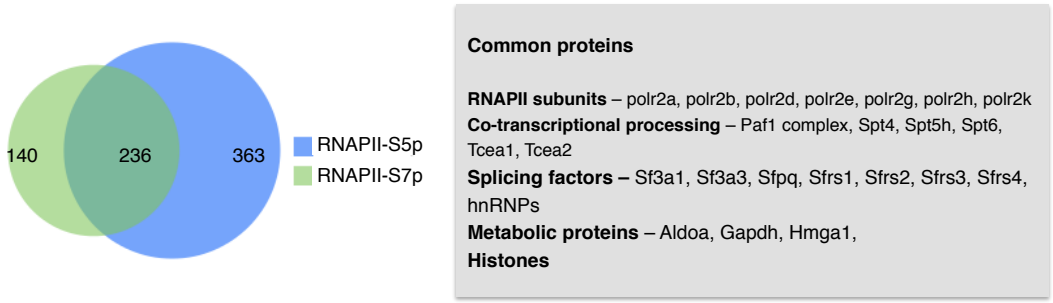


Figure 3.14. pChIP comparison between RNAPII-S5p and RNAPII-S7p pChIP. (A) Venn diagram representing the proteins found common or selectively in pChIP experiments. Examples of proteins common in both RNAPII-S5p and -S7p.

These analyses highlight the feasibility of pChIP in detecting significant numbers of proteins specifically associated with RNAPII dependent processes, such as transcription itself or RNA processing which mostly occurs co-transcriptionally. However, label-free MS analyses are mostly semi-quantitative, not allowing the assessment of whether a protein that may be commonly identified in two different datasets may be differentially enriched in one versus the other. For proteins uniquely associated with one or another

pChIP (S5p or S7p), it is difficult to assess whether these proteins were not identified in one of the samples, e.g. due to technical (low sequencing depth) or actual biological affinity to different RNAPII modifications.

3.5. Discussion

ES cells are self-renewing and pluripotent cells that have a tightly regulated gene expression where chromatin forms the platform for TFs, chromatin modifiers and regulatory modules to cascade a range of downstream processes (Jaenisch and Young 2008; Melcer and Meshorer 2010). The diversity of mES cells chromatin can be understood at different levels owing to the resolution of chromatin (DNA and proteins) obtained by different chromatin preparations (Solomon *et al.* 1988; O'Neill and Turner 2003; Zeng *et al.* 2006). I first compared the DNA obtained from the three different chromatin preparations (native, crosslinked and double-crosslinked; Fig. 3.1). After optimising conditions for reverse crosslinking of chromatin, I measured observed the diversity of chromatin proteome from the three preparations by MS (and Coomassie staining) and further compared with published dataset to show specificity and robustness of our MS proteins detection (Fig. 3.2, 3.3, 3.4 and 3.5). I choose fixed chromatin as starting material for subsequent experiments, owing to diversity and protein abundance. Next, I reproduced the DNA-ChIP results for RNAPII modifications at a panel of genes and single gene locus (Fig. 3.4, 3.5 and 3.6). Next I designed and optimised pChIP protocol, experimental and MS conditions (in collaboration with Bram Snijders). Finally I performed RNAPII-S5p and -S7p pChIP identifying range of known and novel protein interactions including independent confirmation for Polycomb proteins association with RNAPII-S5p on chromatin (Fig. 3.12), previously only identified for a small number of genomic regions by sequential ChIP (Brookes *et al.* 2012).

3.5.1. Chromatin diversity captured by different chromatin preparations.

The resolution of native chromatin spans mono and di-nucleosomes that corresponds to 150-400 base pairs (bp) of DNA (O'Neill and Turner 2003). Active gene promoters have open chromatin characterised by spaced nucleosomes and at the resolution of native ChIP, only histones, histone modifications and strongly bound transcription factors can be studied. In fixed nuclear chromatin (fixed nuclear extract) formaldehyde crosslinks the interactions (DNA-protein, protein-protein and RNA-protein). Upon sonication we observe range of chromatin (DNA) fragments that span from 300bp - 1.2kbp representing larger fragments with crosslinked interactions. Fixed ChIP is therefore used to observe occupancy of transcription factors, enzymes and proteins that interact on chromatin and cascade downstream processes. In our intermediate depth MS analysis of chromatin proteome at intermediate depth presented here, we observed enrichment of 75% more proteins (978) in fixed chromatin (2298) compared to native chromatin (1320 proteins) consistent with resolution and nature of chromatin preparation. The diversity of native chromatin proteome is evident with the enrichment for pluripotency factors, G-protein signalling and RNA processing factors suggesting strong and intricate link with chromatin (nucleosomes). Enrichment for Oct4, Nanog and other pluripotency factors in our limited depth native chromatin run suggests a closer and tightknit regulation of stem cell transcription factors on chromatin while composition of fixed chromatin proteome was diverse (transcriptional machinery, splicing factors, transport and metabolic proteins) consistent with composition of nuclear proteome.

The ability to detect proteins from fixed chromatin preparation also depends upon our ability to efficiently reverse crosslink, elute and extract proteins. This step is quite critical, as un-crosslinked proteins tend to aggregate and not pass through the SDS-PAGE gel. In addition, MS allows identification of proteins present in these aggregates as peptides are eluted to allow protein identification. We observe that conditions for reverse crosslinking of chromatin

proteins (60°C o/n and 100°C for 10min) are optimal for most proteins tested by western blotting. However different proteins might have other optimal reverse crosslinking conditions and might therefore have lower identification in the dataset studied.

3.5.2. Pattern of RNAPII modifications at Active and PRC-repressed genes

RNAPII at PRC-repressed genes contain high levels of S5p but no detectable levels of S7p or S2p (Fig. 3.6) in spite of detection of S5p into coding regions. The levels of S5p are often higher in PRC-repressed genes compared to active genes (Fig. 3.6) while the levels of total RNAPII are similar between active and PRC-repressed genes (Stock *et al.* 2007b). Presence of S5p (in the absence of S7p and S2p) along with Polycomb proteins at coding regions of PRC-repressed genes is consistent with transcription without any expression and suggests a novel mechanism of RNAPII-S5p and Polycomb interplay at these genes.

In addition the RNAPII antibodies used are extensively characterised and detect RNAPII occupancy across a single genes by qPCR and also genome-wide (Brookes *et al.* 2012). However certain commercial antibodies (e.g. against RNAPII-S5p; Abcam ab5408) only detect RNAPII at promoter regions thereby affecting the coverage of proteins on DNA. For pChIP, variations like these may have strong effects. In addition my DNA-ChIP results are robust and consistent with our published work adding confidence to pChIP method.

While ChIP-Sequencing (ChIP-Seq) measures the occupancy of a single protein on the chromatin genome-wide, it does not offer any information on other proteins that co-associate along with it on chromatin. Multiple ChIP-Seq can be performed to identify potential co-associations however this approach is very expensive and requires previous knowledge on proteins. Proteome-ChIP on the other hand is a unbiased approach that dissects the chromatin

bound proteome associated with protein of interest. Further identifying cohorts of proteins that co-associate on chromatin thereby removing the need for hundreds of ChIP-Seq experiments. pChIP with MS provides global information on proteins co-associating on chromatin. Proteome-ChIP differs from conventional peptide pull-down and immunoprecipitations as we use *in-vivo* chromatin extract to enrich for proteins association (rather than nuclear extract or whole cell extract).

After immunoprecipitation, chromatin samples after pChIP are re-suspended in custom Laemmli buffer to effectively and efficiently reverse crosslink and denature chromatin-bound proteins allowing for peptide elution and detection by MS. One disadvantage is that quantification of proteins cannot be done while in Laemmli buffer, which precludes their analysis without separation by SDS-PAGE and in addition efficiency of reversal of crosslinks cannot be calculated (or normalised) between different experimental runs. Therefore I have always performed DNA-ChIP in parallel to pChIP and used the DNA yields after immunoprecipitation as test of robustness. pChIP proteins can be visualised by western blotting, however that severely restricts the applicability of pChIP due to the sample loss during SDS-PAGE (and transfer onto nitrocellulose membrane) and much lower detection limit for proteins in western blot varies from 1-10ng. pChIP and MS analysis is much more sensitive detecting peptides instead of intact proteins, in addition MS sample processing steps (de-staining, alkylation, trypsinisation, chromatography) further remove salts, DNA, RNA, aiding the better separation and protein identification. The major advantage of pChIP and MS analysis is that one run provides identification on several co-associating proteins whereas pChIP with western blot can only provide information on individual protein. Another critical aspect of MS sample processing is the peptide sequencing depth. Complete sequencing depth is achieved when at least one peptide is detected for all the proteins present in the sample. However this equates to abundance of proteins in sample and also chance for MS to capture the peptide spectra for

tandem MS/MS analysis and unreasonably large MS run times. The choice of MS parameters during different runs depends therefore on the sample complexity, efficient chromatographic separation and important reasonable MS run times.

3.5.3. RNAPII Proteome-ChIP

We have demonstrated that pChIP is robust and specifically enriches proteins from input chromatin (Fig. 3.12 and Fig. 3.13). RNAPII-S5p peaks at promoter of active genes and is also present throughout coding regions. In RNAPII-S5p pChIP, we identify >700 proteins involved in range of diverse cellular processes. We identify many components of transcriptional machinery (General transcription factors, Splicing, RNA processing, Paf1 complex) associating with RNAPII-S5p consistent with role in transcription. In addition identify several metabolic proteins associating with S5p. Identification of Polycomb proteins exclusively associating with RNAPII-S5p is an independent validation for RNAPII-Polycomb interplay. RNAPII-S7p transitions transcription from initiation to elongation. RNAPII-S7p peaks at promoter with low levels during coding regions, consistently RNAPII-S7p pChIP identifies range of proteins involved in transcriptional initiation and elongation. Remarkably chromatin modifiers and metabolic proteins are identified suggesting role of metabolic protein during transcription. This diversity of proteins identified with S5p and S7p pChIP is remarkably and highlights the importance of RNAPII and its association.

Using pChIP with conventional MS approaches provides only qualitative information of proteins and their associations. Due to the qualitative aspect of these analyses, we cannot compare the proteins identified from RNAPII-S5p and S7p pChIP with each other due to differences in antibody efficiency, reverse crosslinking and elution efficiency. In addition several normalization parameters would be required to equilibrate sample complexity and MS peptide analysis time for different samples. Even after these normalizations,

we cannot quantitatively associate protein binding to one RNAPII modification or the other. For effective comparison, we require a quantitative approach that can resolve protein dependencies to RNAPII modifications while normalizing the effects of antibody efficiency and MS parameters.

4. Optimising SILAC labelling and pChIP-SILAC in mES cells

4.1. Research motivation

To quantitatively investigate the chromatin-bound proteome associated with RNAPII modifications, I investigated and optimised conditions for SILAC labelling of mES cells, further performing quality control experiments to confirm similar behaviour of chromatin, DNA, proteins and pluripotency factors between unlabelled and SILAC cells. I aimed to investigate and quantify the proteomic differences between label-free and SILAC samples including chromatin and pChIP samples. I also aimed to adapt and extend the SILAC method to compare and contrast proteomes associated with different post-translational modifications on single RNAPII molecules (S5p and S7p). Furthermore, I explored experimental strategies to best unravel and quantitatively dissect the chromatin proteome landscape associated with RNAPII modifications (S5p and S7p).

All of the the MS experiments were carried out in collaboration with Dr. Bram Snijders at Proteomics facility at MRC-CSC. Dr. Snijders also helped with sample pre-processing for MS, performed all MS run time operations and quantified MS spectra.

4.2. Culture conditions for mES cells.

Embryonic stem (ES) cells derived from ICM can be proliferated in vitro under appropriate culture conditions (Evans and Kaufman 1981; Martin 1981). These cells have the ability to undergo cell divisions without differentiation so as to produce pluripotent progeny when cultured in vitro (Chambers and Smith 2004) a property referred to as self-renewal, which occurs via symmetrical cell division. ES cells can be propagated under culture conditions in the presence of serum and in co-culture with a layer of fibroblasts (feeder layers) (Chambers and Smith 2004). These feeders act by producing a signal that inhibits ES cell differentiation (Smith and Hooper 1983). Subsequent studies of the conditioned medium identified the active component as leukaemia

inhibitory factor (LIF) (Smith *et al.* 1988; Williams *et al.* 1988). This finding indicated that LIF supplementation along with serum provides all necessary factors required for ES cell proliferation and maintenance. Other important signalling molecules, cytokines, extrinsic regulators and factors that contribute to ES cell self-renewal include, BMP4, BMP2 or GDF6, Wnt signalling, glycogen synthase kinase-3 (GSK-3) during clonal propagation of ES cells and during their *de novo* derivation (Ying *et al.* 2003; Chambers and Smith 2004) The composition of media is also contains necessary growth factors that allow master regulators (Oct4, Sox2 and Nanog) to mediate transcriptional and protein interactional network for pluripotency in ES cells (Wang *et al.* 2006; Orkin *et al.* 2008; van den Berg *et al.* 2010).

ES cells have a tightly regulated transcriptome and proteome that facilitates the need to self-renew, remain pluripotent and respond to appropriate differentiation cues. Proteomic studies have identified additional proteins and factors that are essential for mES cells and have further shed light on the tightly regulated protein network (Unwin *et al.* 2003; Prudhomme *et al.* 2004; Gesslbauer *et al.* 2006; Van Hoof *et al.* 2006; Graumann *et al.* 2008).

4.3. SILAC labelling and MS quantification of proteins.

Stable isotope labelling of amino acids in cell culture (SILAC) offers a simple and practical approach to perform quantitative proteomics (Ong *et al.* 2002). Briefly, essential amino acids (L-lysine and L-arginine) are replaced by their stable heavier isotopes and upon comparison with unlabelled cells, mass spectrometry (MS) efficiently distinguishes the heavier and lighter isotopes leading to quantification of proteomic differences between heavy and light sample. SILAC labelling methods have been further developed and applied to generate entire fly and mouse. Additionally SILAC approaches have also been applied to distinguish proteomic differences between control and disease samples (Mann 2006; Sury *et al.* 2010; Zanivan *et al.* 2012). SILAC MS methods and analysis are more expensive than conventional MS due to the media compositions and requirement of replicate samples (heavier and lighter

amino acids media). Further optimisation of MS conditions and run time are extremely significant and affect depth of proteome (Chen 2008). Due to the nature of the SILAC experiment, robust and quantitative information can be obtained without the need for multiple replicates.

Our lab has previously identified distinct combinations of RNAPII modifications at important developmental regulator genes in mES cells that are silenced by Polycomb repressive complexes (PRC) to prevent aberrant activation of differentiation programs, but also associate with hallmarks of active chromatin (Brookes and Pombo 2009b; Brookes *et al.* 2012). We have also previously demonstrated the specificity of RNAPII antibodies using specific kinase inhibitors, phosphatase treatments (Xie et al 2006, Stock et al 2007), and peptide assays (Brookes et al 2012; data not shown from H. Kimura and D. Eick laboratories). Distinctive genome-wide occupancy of RNAPII modifications was identified at the subgroup of genes associated with Polycomb repression, by ChIP-Sequencing (and by ChIP-qPCR at subset of genes) (Brookes *et al.* 2012). In particular with an interest in identifying proteins associated with RNAPII-S5p associated at Polycomb repressed chromatin, I have worked to develop an unbiased and quantitative assay to unravel the proteome associated with RNAPII-bound chromatin.

4.4. Results

4.4.1. SILAC amino acid concentration is important for mES cells viability

Standard conditions for culture of mES cells includes media (GMEM/DMEM - high glucose) supplemented with fetal calf or fetal bovine serum (FCS/FBS) and additional factors (L-glutamine, sodium pyruvate, non-essential amino acids, LIF and β-mercaptoethanol). Cells are grown at 37°C with 5% CO_2 and in 95% humidity. The composition of FCS/FBS is highly undefined, containing a range of growth factors, proteins, chemokines, vitamins, minerals and complex lipids, which supplement the signalling, metabolism, and growth of mES cells. For SILAC labelling of mES cells, it was essential to use a defined culture medium wherein the concentration of essential amino acids (lysine, arginine and their heavier isotopes) can be appropriately added to support mES cell growth and maintained without affecting mES cell characteristics.

Initially, I optimised the SILAC media and culture conditions for OS25-mES cells. Using SILAC-DMEM media (lacking lysine and arginine amino acids), I tried supplementing media with different concentrations of L-arginine (240, 87.2 and 84 mg/L) and L-lysine (40, 152.8 and 146 mg/L), as reported in literature (Blagoev and Mann 2006; Bendall *et al.* 2008), along with the use of dialysed FCS or serum replacement media (Knockout serum replacement; KOSR). For OS25-mES cells, SILAC-DMEM supplemented with KOSR, L-arginine (84 mg/L, or 0.398 mM) and L-lysine (146 mg/L, or 0.798 mM) were optimal. Other concentrations of SILAC amino acids and conditions either led to massive cell death or to differentiation (assessed by visual inspection of cell morphology). Table 4.1 lists the conditions found optimal for SILAC labelling of OS25-mES cells; they have also been successfully used for growing other mES cells lines, E14-mES and 46C-mES, in our laboratory.

Table 4.1 Media composition for SILAC labelling of mES cells. List of components added to SILAC-DMEM media along with concentrations of heavy and light isotopes of essential amino acids.

SILAC media composition	
Media components	**Final volume/concentration**
SILAC-DMEM	500 ml
Knockout serum replacement	15 %
β-mercaptoethanol	0.1 mM
L-glutamine	2 mM
Sodium pyruvate	1 mM
Non-essential amino acids	1%
Leukemia inhibitory factor	1000 U/ml
Hygromycin	0.1 mg/ml
SILAC amino acids	
Heavy isotope	
L-Lysine HCl ($^{13}C_6\ H_{14}\ ^{15}N_2\ O_2$)	0.798 mM
L-Arginine HCl ($^{13}C_6\ H_{14}\ ^{15}N_4\ O_2$)	0.398 mM
Light istope	
L-Lysine HCl ($^{12}C_6\ H_{14}\ ^{14}N_2\ O_{2)}$	0.798 mM
L-Arginine HCl ($^{12}C_6\ H_{14}{}^{14}N_4\ O_2$)	0.398 mM

4.4.2. Culturing mES-OS25 cells in SILAC conditions does not affect pluripotency markers

Before using the SILAC-labelled mES cells for quantitative pChIP, I first performed different quality control assays to check whether the SILAC labelling affected mES cell characteristics (Fig. 4.1). Using alkaline phosphatase (AP) staining of OS25-mES cells grown in normal or SILAC conditions, I confirmed that cells acquired similar levels of AP staining in SILAC as in normal conditions, confirming their stemness or pluripotency potential (Fig. 4.1A). I also measured RNA levels of three genes fundamental for mES cell pluripotency and self-renewal (Oct4, Sox2 and Nanog), and I observed that similar levels of RNA were expressed in SILAC as in normal conditions (Fig. 4.1B). Next, I compared overall protein composition by Coomassie staining of whole cell extract proteins and observed overall similar distribution (Fig. 4.1C). Assessing protein levels of pluripotency genes (Oct4, Sox2 and Nanog) by western blotting showed similar protein amounts in normal and SILAC conditions (Fig. 4.1D). Finally, I checked the distribution of RNAPII-S5p and Nanog in normal and SILAC cells by 3D-

immunofluorescence. RNAPII-S5p is distributed throughout the nucleoplasm, excluding sites of heterochromatin (Xie and Pombo 2006; Silva *et al.* 2009). Nanog protein is known to fluctuate its expression in mES cells (Silva *et al.* 2009) and fluctuating Nanog levels can be observed in normal and SILAC-labelled cells.

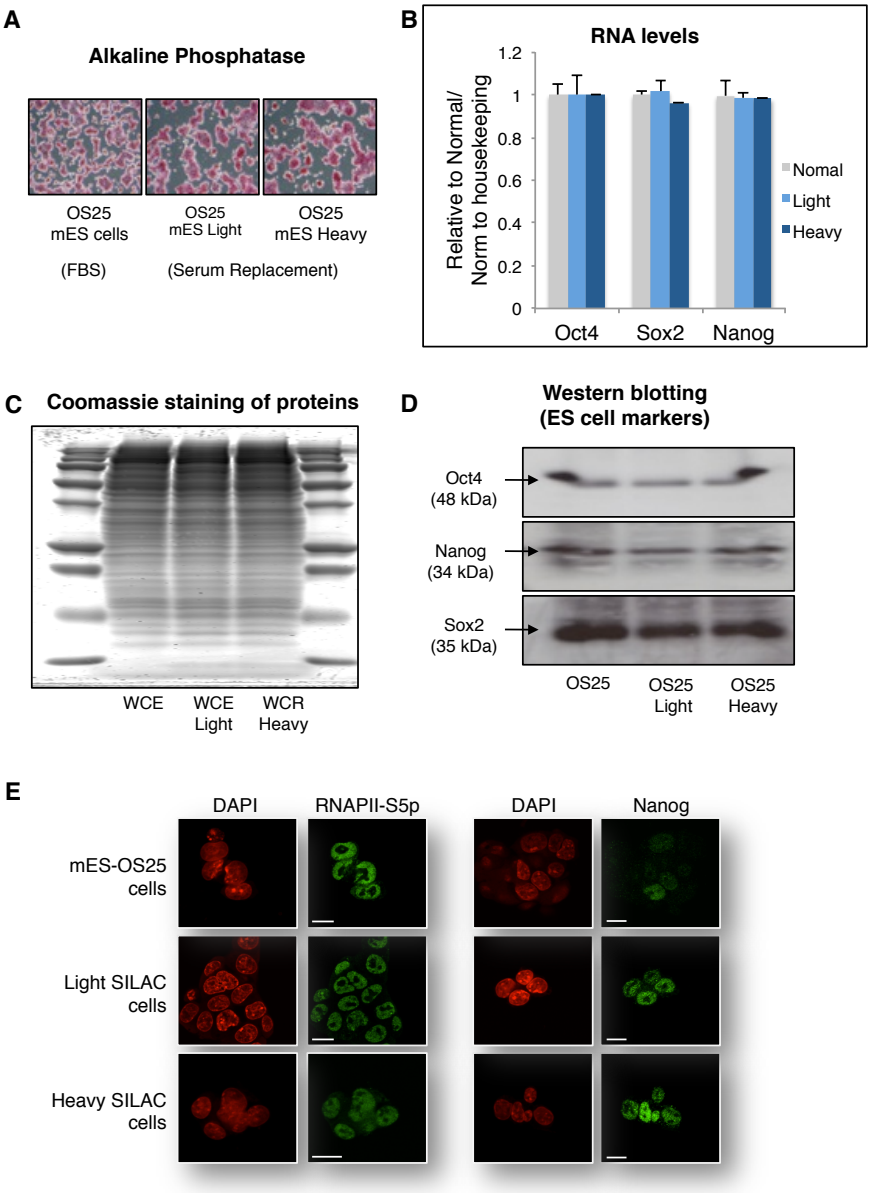


Figure 4.1 mES cell characteristics are not affected after SILAC labelling. (A) Alkaline phosphatase staining of normal mES cells (FBS cultured) and in SILAC conditions. (B) Measuring RNA levels of pluripotency genes Oct4, Sox2 and Nanog. RNA levels for normal and SILAC conditions were first normalised to expression of housekeeping genes (βactin, Ubiquitin-c, Glucose-6-phosphate dehydrogenase),

before normalizing SILAC samples to level of RNA produced in normal conditions. (C) Coomassie staining shows similar distribution of proteins in the whole cell extract (WCE) of normal and SILAC cells. (D) Western blotting confirms presence of pluripotency markers in all samples. (E) Immunofluorescence for RNAPII-S5p and Nanog protein shows similar pattern of protein distribution across cells in normal and SILAC conditions.

4.4.3. Verifying incorporation of heavy and light amino acids in whole cell extract by MS

After confirming the pluripotency markers of mES cells grown in SILAC conditions, I next checked the incorporation of lysine and arginine isotopes in the proteins in cells grown in SILAC conditions. Whole cell extracts (WCE) were prepared from SILAC cells (minimum 4-6 passages, equivalent to 8-12 doubling times), mixed 1:1 (according to protein concentration determined by Bradford assay), before separation by SDS-PAGE and staining with Coomassie. One or two fractions per gel were analysed using low-depth (shorter) MS runs, which resulted in the identification of 170-392 proteins (Fig. 4.2). We observed >97% labelling efficiency; only light peptides were detected in light samples and heavy peptides in heavy samples. Common MS contaminants (keratins, trypsin, control peptides, etc.) were identified and filtered, as they are consistently enriched only in light peptides. Examples of mass spectra of lysine and arginine containing heavy peptides are shown in Fig. 4.2. We also did not detect any proline conversion, a difficulty reported in the literature in different mES cell lines (Van Hoof *et al.* 2007).

	Light only sample	Heavy only sample	Heavy and Light mixed sample
Number of gel slices analyzed	2	1	1
Total number of proteins identified	379	190	170

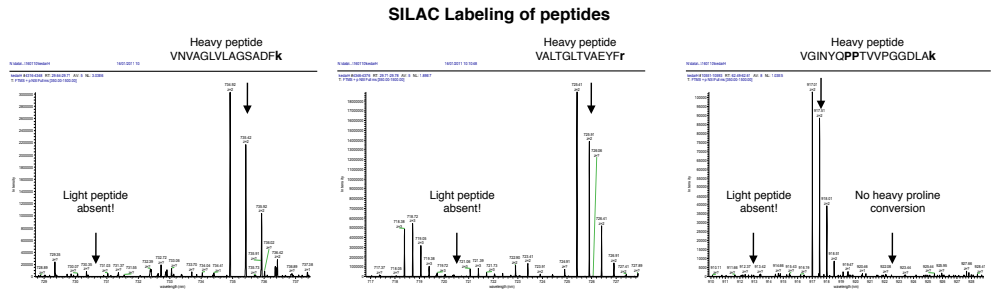


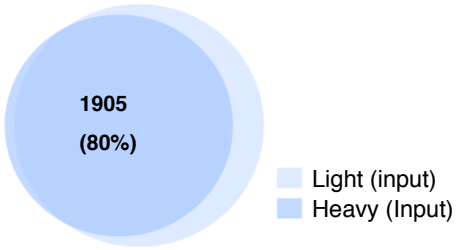
Figure 4.2 SILAC labelling efficiency measured by MS. Whole cell extracts were made from SILAC heavy and light labelled cells (minimum 4-6 passages) and analysed by MS.

Only heavy peptides were detected in the heavy-only sample, and only light peptides were detected in light-only samples. A few examples of mass spectra from heavy WCE samples after MS analysis and database searching.

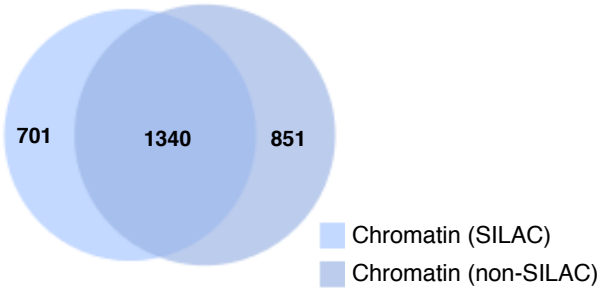
4.4.4. MS analysis on SILAC chromatin

SILAC labelling allows the comparison of the relative amount of proteins present in a pair of samples. Given that our input chromatin is a complex MS sample due to the efficiency of reverse crosslinking, I tested how robustly could proteins be detected in the SILAC input chromatin, in comparison with input chromatin from mES cells grown in normal conditions (as in Fig. 3.12). Chromatin (heavy) was mixed with chromatin (light) in 1:1 ratio (by chromatin concentration) and analysed by MS in low-depth run. We identified ~2400 proteins in input chromatin samples; >80% of these proteins (1905 proteins) had robust SILAC ratios (ratios 1.0±0.2; Fig. 4.3). Comparing SILAC with non-SILAC input chromatin, there was large overlap of proteins common in both datasets (1340 proteins; Fig. 4.3B). Over 20 significant GO terms are enriched and shared between all the datasets (Fig. 4.3 C). Taken together, these results suggest similar chromatin composition the non-SILAC and SILAC chromatin, with an expected variation in number of proteins due to MS peptide sequencing depth.

A



B



C

Significant GO terms (SILAC and non SILAC chromatin)

GO terms (# of proteins)	SILAC	Overlap	Non-SILAC
Translation (371)	46 %	38 %	48 %
mRNA processing (286)	47 %	32 %	49 %
RNA splicing (228)	41 %	36 %	52 %
Protein folding (141)	39 %	34 %	53 %

* p-value < $10e^{-5}$

Figure 4.3 MS analysis of SILAC input chromatin (heavy/light). Heavy and light input chromatin was mixed in 1:1 ratio as per chromatin concentration (alkaline lysis as described in section 2.2.1.1). (A) In low depth MS run, more than 80% of chromatin proteins were identified and quantified (ratio 1.0±0.2) (B) Comparison between SILAC and non-SILAC chromatin (as in Fig. 3.4). (C) Examples of significant GO terms in SILAC chromatin, overlap and in non-SILAC chromatin. More than 20 GO terms were enriched in all three groups. Percentages represent the proportion of identified proteins relative to the total number of annotated proteins in each GO category.

4.4.5. RNAPII occupancy in SILAC mES chromatin

To test whether the chromatin occupancy of RNAPII modifications was not affected, I performed DNA-ChIP on SILAC-labelled light and heavy chromatin, before performing pChIP with SILAC-labelled chromatin (Fig. 4.4). As observed in non-SILAC conditions (Fig. 3.7 and (Stock *et al.* 2007b)), RNAPII-S5p was enriched at promoters and coding regions of active and PRC-repressed genes in both Heavy and Light chromatin. RNAPII-S7p was enriched only at promoters and coding regions of active genes and RNAPII-

S2p was also predominantly enriched at coding regions of active genes, as expected. Inactive genes had no enrichment for either RNAPII modification. Mock pChIP was performed in parallel to all pChIP experiments and demonstrates minor levels of non-specific ChIP enrichment.

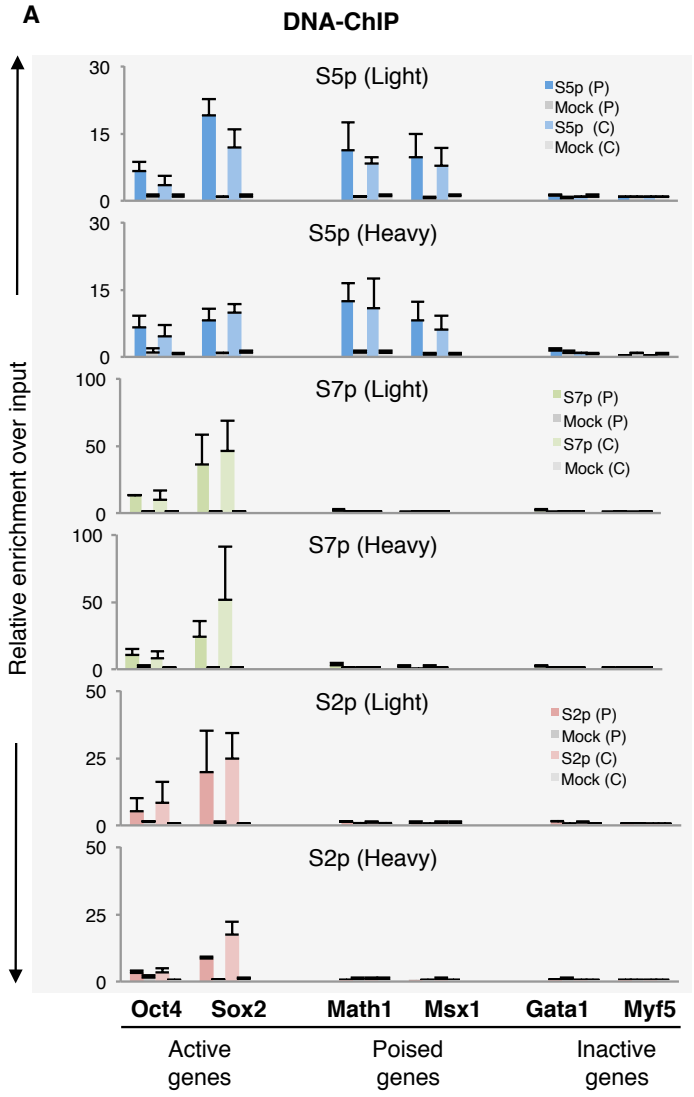


Figure 4.4 RNAPII occupancy in SILAC-labelled mES cells. Occupancy of RNAPII-S5p (dark and light blue), -S7p (dark and light green) and -S2p (dark and light pink) as measured by DNA-ChIP and qRT-PCR at promoters (P) and coding (C) regions of a panel of Active genes (Oct4, Sox2), PRC-repressed genes (Math1, Msx1) and Inactive genes (Gata1, Myf5) in SILAC-labelled light (light colours) and heavy (dark) cells. Dark and light grey bars represent background enrichment levels as measured by ChIP in the absence of primary antibody (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. Mean and standard deviation are representative of 3 independent biological replicates.

4.4.6. Determining the proteome of chromatin occupied by RNAPII-S5p using SILAC pChIP

Having performed several quality controls on SILAC cells and chromatin and confirmed the occupancy of RNAPII modifications, I next performed RNAPII-S5p SILAC pChIP. RNAPII-S5p pChIP was performed on both heavy and light chromatin, before mixing heavy and light ChIP eluates (1:1 by volume) and analysing the extracted proteins by MS. SILAC ratios between heavy and light chromatin were obtained. To normalise for differences in the efficiency of protein elution and allow effective comparison of protein enrichment relative to Rpb1, all SILAC ratios were normalised to the ratio of Rpb1, the subunit immunoprecipitated. For example, in an initial low-depth MS run, I first observed that pChIP ratio for RNAPII subunit Rpb1 was enriched in heavy pChIP relative to light pChIP. This was mainly due to a slightly larger starting volume in heavy pChIP (700 µg of chromatin – 750 µl) than in the light pChIP (700 µg of chromatin – 630 µl), but will also depend on sample complexity when contrasting pChIP samples using different antibodies. Even at this relatively low depth, we could identify 309 proteins robustly with SILAC pChIP ratios (1.0±0.2), *i.e.* detected in light and heavy peptides with similar efficiency, including robust identification of six RNAPII subunits, three Polycomb proteins and, interestingly, also components of the DNA replication machinery (Fig. 4.5).

I asked next whether proteins identified in RNAPII-S5p SILAC pChIP were similar to proteins identified in non-SILAC pChIP experiments and what GO terms were enriched. The number of proteins common to both datasets was 133, with a further 176 proteins specifically identified in RNAPII-S5p SILAC pChIP (Fig. 4.5), which could be explained by limited depth, as previously observed in the comparisons between SILAC and non-SILAC input chromatin (section 4.3.4). GO analyses of RNAPII-S5p pChIP identified common terms in the SILAC and non-SILAC datasets, such as RNA splicing and mRNA processing. The relatively small overlap in these preliminary tests suggests significant complexity of the proteome of chromatin bound by RNAPII and the

importance of running the MS analyses at higher depth. GO analyses and scanning the datasets for specific proteins, such as RNAPII subunits and other RNAPII-associated proteins such as splicing factors, suggest that SILAC pChIP can specifically enrich and robustly identify proteins associated with chromatin occupied by RNAPII complexes characterized by different post-translational modifications.

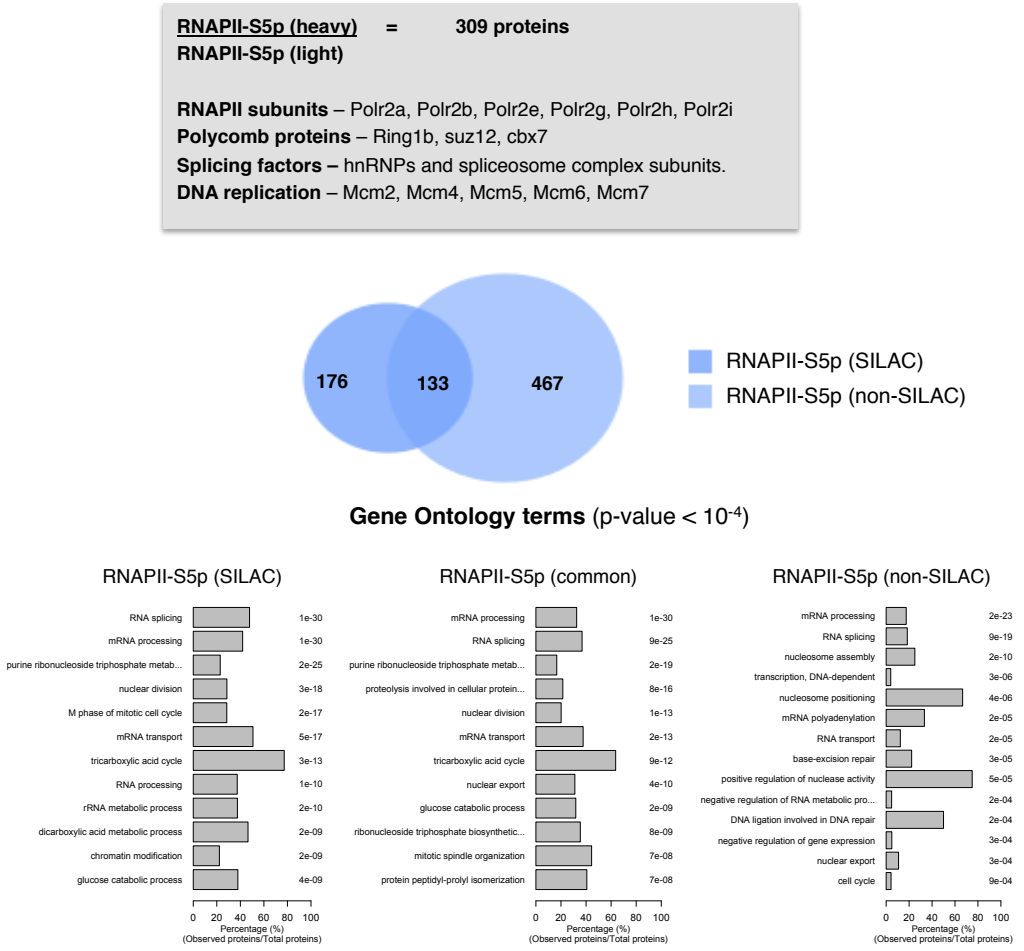


Figure 4.5 MS analysis of RNAPII-S5p SILAC-pChIP. Heavy and light RNAPII-S5p pChIP eluates were mixed (equal volume) and analysed by MS. (A) Low-depth MS run identified over 300 proteins including RNAPII subunits and other interesting proteins. (B) Comparison between RNAPII-S5p SILAC and non-SILAC pChIP (as in Fig. 3.12) represented by a Venn diagram. (C) Enriched GO terms for proteins shared and specific between SILAC and non-SILAC pChIP.

4.4.7. Comparison between heavy RNAPII-S5p pChIP and light RNAPII-S7p pChIP

To begin comparing the proteome associated with different forms of RNAPII modifications, I next performed SILAC pChIP experiments using antibodies for RNAPII-S5p and -S7p. In this experiment, I performed RNAPII-S5p pChIP on heavy chromatin and RNAPII-S7p pChIP on light chromatin. Protein eluates were mixed (equal volume) and analysed by MS producing a dataset with 593 proteins. After normalization of SILAC ratios (Heavy/Light) to Rpb1 ratio (immunoprecipitated with both antibodies), we classified proteins according to whether they were more abundantly detected in association with one of the RNAPII modifications (S5p or S7p) if their ratios were >1.2 (for S5p; 382 proteins) or <0.8 (for S7p; 109 proteins). Proteins enriched for S5p included chromobox proteins (involved with silencing and heterochromatin formation), DNA replication, DNA methylation and a protein essential for mES cell pluripotency (Fig. 4.6B). Proteins enriched for S7p have functional roles in transcription initiation and elongation. Common proteins (0.8 < ratio < 1.2) were enriched for RNAPII subunits, histones and components of co-transcriptional processing, the latter consistent with the fact that RNAPII-S5p and S7p are detected in coding regions. GO analyses (Fig. 4.6) show enrichment for gene expression and mRNA related terms in both RNAPII-S5p and S7p datasets and specific enrichment for promoter terms in S7p.

These preliminary pChIP analyses highlight the power of pChIP to capture dynamics of protein recruitment to chromatin during transcriptional processes. A more thorough analysis, that would include pChIP for RNAPII-S2p (hallmark of transcription elongation), would be necessary to distinguish components of transcriptional elongation from transcriptional initiation to allow finer unravelling of the chromatin communities associated with RNAPII during the sequence of co-transcriptional processes that take place from initiation to termination. Together with the preliminary analyses shown here, that put in evidence the importance of MS depth, it was also clear that the inclusion of replicate pChIP eluates, in both Heavy and Light chromatin (forward and

reverse SILAC experiments), for each antibody, would aid in identifying protein enrichments more robustly and in classifying their dependency with RNAPII modification on S5p, S7p or S2p.

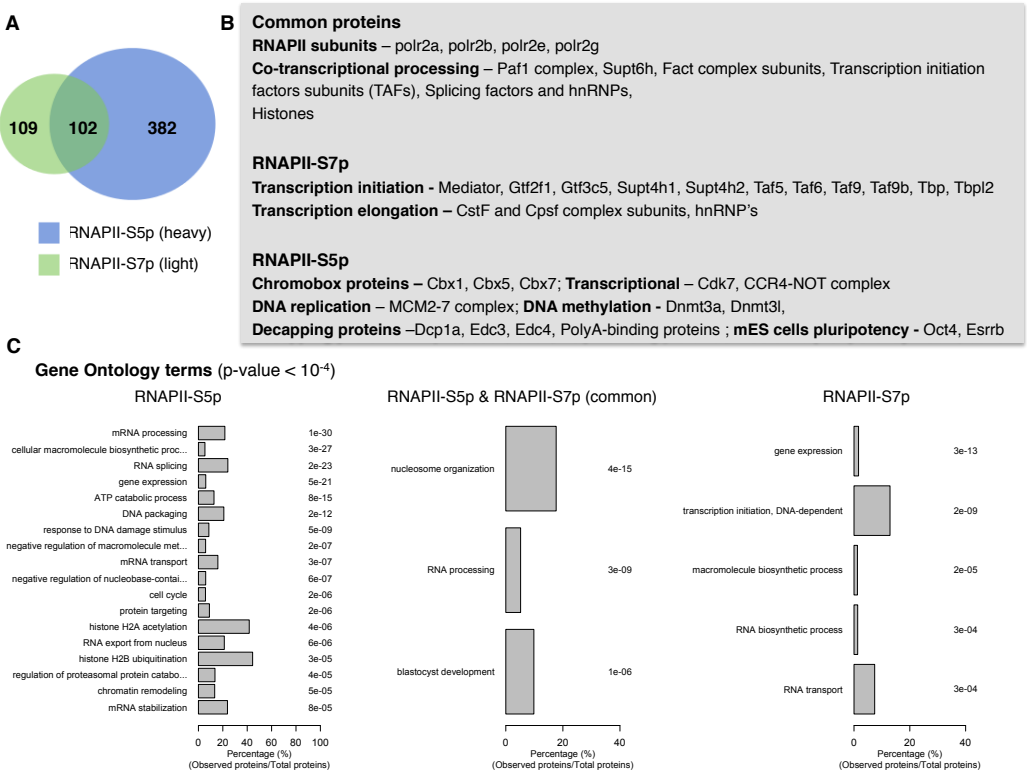


Figure 4.6 MS analysis of SILAC pChIP for RNAPII-S5p (Heavy) and RNAPII-S7p (Light). RNAPII-S5p (heavy) pChIP was mixed with RNAPII-S7p (light) pChIP (by volume) and analysed by MS. (A) Proteins enriched for RNAPII-S5p (ratio > 1.2), enriched for RNAPII-S7p (ratio < 0.8) and common (0.8 < ratio < 1.2) are represented in a Venn diagram. (B) Examples of proteins found in the different categories. (C) Significant GO terms enriched after pChIP in either the specific or shared RNAPII modifications.

4.5. Discussion

4.5.1. SILAC labelling retains mES cell characteristics

We have shown that media composition for culturing mES cells is highly undefined and contains a variety of growth factors, cytokines, signalling molecules in addition to other unknown components of serum. mES cells are highly sensitive to their environment, media composition and insufficient stimulus can lead from mild effects, including compromised energy metabolism to more severe effects including differentiation and cell death

(Brinster and Harstad 1977; Gassmann *et al.* 1996; Follmar *et al.* 2006). I tested and optimised the conditions for SILAC labelling of our OS25-mES cells. Remarkably, SILAC composition of amino acids between different mES cells (and between other cell types) can vary extensively and have to be optimised in each case to avoid mild and severe effects.

Broad comparisons of SILAC and unlabelled mES cells showed (a) similar levels of alkaline phosphatase (Fig. 4.1), a marker for pluripotency, (b) RNA and protein levels of master regulator genes (Fig. 4.1; Oct4, Sox2 and Nanog) and overall similar protein size profile (Fig. 4.1; Coomassie staining). Indirect immunofluorescence detection of RNAPII-S5p and Nanog proteins extends our observation that SILAC labelling does not affect mES cell characteristics. However, we have not performed a global genome, transcriptome and whole cell proteome analysis that might help negate any effect of SILAC labelling.

4.5.2. Advantages and pitfalls of MS analysis on complex samples.

The most important pre-requisite for SILAC is the identification of peptides in both the heavy and light samples, which can be challenging when studying complex samples, in particular for the least abundant proteins. However, SILAC approach is more robust, sensitive and specific towards identifying proteins than of methods, and drastically reduces the need for large number of replicates. In addition, protein identification with conventional and SILAC MS is also affected by PTMs of proteins. PTMs affect the migration of peptides during chromatography and alter the charge state of the peptides during MS analyses, thereby making the acquired spectra incomparable to reference database. In addition, MS works sequentially therefore only a few peptides (of the total peptides pool) are analysed for tandem MS/MS. The combination of PTM's and sequential peptide analysis adversely affects the individual peptides corresponding to a single protein, in particular for SILAC quantification wherein peptides have to be identified in both heavy and light fractions within finite MS run time. Peptide sequencing depth is also interlinked with MS run time and complete sequencing depth would require

large amounts of starting material and unreasonable MS run times. The SILAC ratio for a given protein is computed as a median of the individual peptide ratios calculated from the spectra of peptides identified in both heavy and light. Many peptides are detected only in heavy or light version, and these peptides have often quite variable ratios.

Our MS analysis of chromatin from SILAC-labelled heavy and light cells is a low-depth MS run allowed us to robustly identify ~1900 proteins in both heavy and light chromatin and with consistent SILAC ratios. Our threshold for comparing heavy and light SILAC ratios was relatively stringent ($0.8 < Ratio < 1.2$) and was performed without any correction for volume or protein concentration. The inconsistency of the additional 381 proteins identified in Fig. 4.3A is attributed to the above MS issues. Comparisons of proteins identified in SILAC and unlabelled chromatin (Fig. 4.3B; 851 versus 701 proteins) are also in the range expected for comparisons of equivalent samples, and these differences are attributed to limited sequencing depth, different MS run times and/or partial detection in SILAC samples.

4.5.3. Quantitative analysis of RNAPII-S5p and S7p proteome.

We have shown that RNAPII-pChIP specifically and robustly identifies chromatin-bound proteome over input chromatin. From low depth MS analysis of RNAPII-S5p pChIP done on SILAC chromatin, we identified 300 proteins with our stringent threshold ($0.8 < Ratio < 1.2$) after normalization to Rpb1 pChIP-SILAC ratio. Even in the preliminary low depth run, we could remarkably observe proteins involved similar functions including transcription machinery and Polycomb proteins. Comparison SILAC RNAPII-S5p pChIP with label-free RNAPII-S5p pChIP yields enrichment of similar GO processes and extends our observation.

Our analysis of SILAC pChIP with S5p and S7p RNAPII modifications (S5p/S7p; Fig. 4.6) highlights our ability to dissect the chromatin bound proteome and unravel specific protein communities bound to chromatin

occupied by RNAPII complexes with different levels of S5 or S7 phosphorylation. Contrasting S5p with S7p pChIP, we robustly identify transcriptional apparatus (RNAPII subunits and co-transcriptional machinery) shared between both the marks, consistent with our genome-wide distribution and DNA-ChIP results (Stock *et al.* 2007b; Brookes *et al.* 2012). In spite of a lower number of proteins identified in the S7p pChIP, the proteins identified have roles in the transition between transcription initiation and elongation at active genes, consistent with the current literature of S7p interplay with S5p and S2p. Remarkably, S5p-only proteins (i.e. proteins associating on chromatin not at actively transcribing genes) are enriched for various non-transcriptional processes. We identify Polycomb proteins exclusively interacting with RNAPII-S5p, confirming but going beyond limited analyses of RNAPII-S5p and Polycomb co-association on chromatin at a small number of genomic regions (Stock *et al.* 2007b; Brookes *et al.* 2012). The identification of several additional processes associated with S5p, including replication, DNA methylation, pluripotency and cell cycle, further elucidate the importance of RNAPII roles on chromatin beyond transcription itself and the novelty of pChIP to identify these interactions in an unbiased manner.

Using pChIP on just two modifications on the RNAPII-CTD, we have been able to capture and dissect the chromatin-bound proteome further unravelling the landscape associated at actively transcribing genes (proteins common to S5p and S7p and S7p-only proteins) and at other chromatin regions including PRC-repressed genes (S5p only). To further unravel the RNAPII landscape, we require additional hallmarks of transcription stages (S2p; transcriptional elongation) and a comprehensive experimental scheme with redundant pChIP runs to avoid the need for replicates and robustly dissect protein dependencies to RNAPII modifications.

5. Unravelling the interactome in mES cells using pChIP

5.1. Research motivation

To investigate the chromatin processes that co-associate distinctly with specific RNAPII modifications on chromatin, we developed a novel unbiased comprehensive strategy involving 16 different SILAC pChIP samples for four kinds of immunoprecipitation (S5p, S7p, S2p and mock) using light or heavy chromatin, which were combined in 12 ways.

To characterise the protein dependencies to specific RNAPII modifications identified across SILAC pChIP datasets that compare pChIP experiments using the four different immunoprecipitations, I devised a data analysis strategy to robustly identify the proteome associated with chromatin-bound RNAPII. In particular, I was interested in investigating and robustly identifying the chromatin processes that co-associate with PRC-repressed RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) in mES cells, by comparing these proteins (and processes) with those associated with RNAPII complexes engaged in productive transcription ($S5p^{+}S7p^{+}S2p^{+}$). I was also interested in identifying other novel processes associated with RNAPII marked by S5p and, in addition, identifying further components of the general transcription machinery and proteins specific to mES cells. Identification of novel chromatin states associated with RNAPII-S5p would shed light on RNAPII interplay in dynamic gene regulatory network in mES cells.

All of the MS experiments were carried out in collaboration with Dr. Bram Snijders at Proteomics facility at MRC-CSC. Dr. Snijders also helped with sample pre-processing for MS, performed all MS run time operations and quantified MS spectra. Borislav Vangelov and Prof. Mauricio Barahona performed the regression analysis (Table 5.2).

5.2. Distinct RNAPII complexes at Polycomb-repressed genes.

RNAPII is present both at actively transcribing genes and at PRC-repressed genes in mES cells. RNAPII transcription at active genes is coupled with expression with Rpb1 modification $S5p^{+}S7p^{+}S2p^{+}$; this configuration of the polymerase leads to the production of stable and mature mRNA that is translated into protein in the cytoplasm. RNAPII transcription at PRC-repressed genes is remarkably different with a distinct Rpb1 conformation ($S5p^{+}S7p^{-}S2p^{-}$) that does not lead to mRNA production (Stock *et al.* 2007b).

Single gene analyses at a panel of active and PRC-repressed genes highlight the distinct differences in chromatin architecture (Stock *et al.* 2007b). Active genes contain open chromatin mark H3K4me3, while PRC-repressed genes are marked by bivalent chromatin, *i.e.* Polycomb instigated repressive H3K27me3 and H2Aub1, but also active H3K4me3. The pattern of individual RNAPII modification at active and PRC-repressed genes is distinct, although with similar total RNAPII occupancy (Stock *et al.* 2007b; Brookes and Pombo 2009b).

Preforming genome-wide ChIP-Seq analysis and correlating with mRNA-See, our laboratory has demonstrated that PRC-repressed genes encompass 25% of mES cell genome with distinct RNAPII conformation ($S5p^{+}S7p^{-}S2p^{-}$) and bivalent histone marks ($H3K4me3^{+}H3K27me3^{+}H2Aub^{+}$). PRC-repressed genes encode for important developmental regulators, metabolic genes and several signalling pathways (Brookes *et al.* 2012). Many of the silenced developmental regulator genes (PRC-repressed) in mES cells are activated upon appropriate differentiation cues and during lineage commitment highlighting the role for RNAPII-Polycomb interplay to balance the need of mES cells stemness and differentiation.

My motivation was to develop pChIP to identify the proteins that simultaneously bind to PRC-repressed chromatin along with RNAPII and

Polycomb proteins. In order to do that, we aimed to robustly contrast the proteins that are involved in active transcription (marked by $S5p^{+}S7p^{+}S2p^{+}$) and differentiate from PRC-repressed chromatin ($S5p^{+}S7p^{-}S2p^{-}$).

5.3. Re-plotting ChIP-Sequencing profiles for most robust active, PRC-repressed and inactive genes

To get a concise picture of the gene expression program coordinated by RNAPII on chromatin, I first started by re-plotting ChIP-Sequencing profiles for most robust active, PRC-repressed and inactive genes generated from our lab for RNAPII modifications in mES cells (Brookes *et al.* 2012). From the published classification, I took a subset of genes (representing core active, PRC-repressed and inactive genes based on RNAPII modifications and Polycomb mediated histone modification - H2Aub) and re-plotted the distribution of RNAPII and histone modifications across average normalised gene start and end sites (-3kb to +3kb). Active genes have high occupancy of all RNAPII modifications (S5p, S7p and S2p) with no H2Aub, while PRC-repressed genes have high S5p and H2Aub with no detectable S7p or S2p. Inactive genes serve as baseline for comparison (Fig. 5.1). Of the different chromatin states present in mES cells, we hypothesised that pChIP using RNAPII antibodies should capture the cohorts of interactors associated with chromatin during transcription at active genes and also distinguish proteins involved in non-transcriptional processes bound to poised forms of RNAPII (for example, Polycomb proteins).

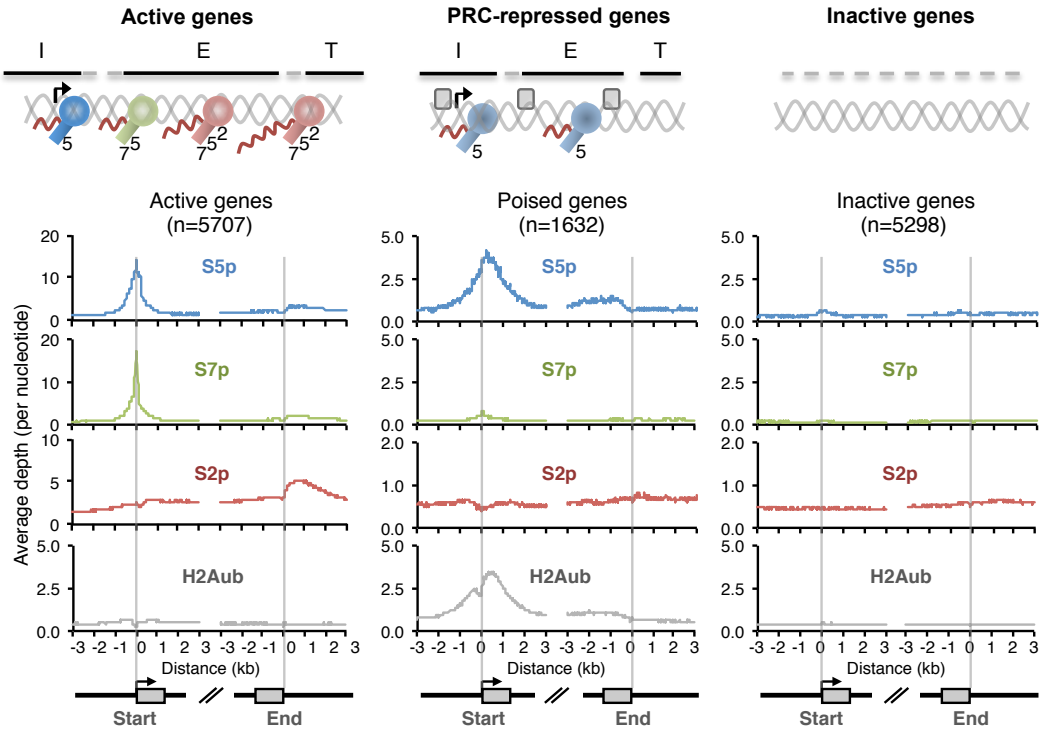


Figure 5.1. Average occupancy of RNAPII modifications in mES cells as mapped by ChIP-Sequencing. ChIP-Seq data from (Brookes *et al.* 2012) was re-plotted for subset of genes (Active, PRC-repressed and Inactive genes) to represent average occupancy of RNAPII modifications (S5p, S7p and S2p) and Polycomb-mediated repressive histone mark (H2Aub) across gene start sites (-3kb to +3kb) and end sites (-3kb to +3kb). Profiles at Inactive genes serve as baseline comparison for enrichment. The y-axis scales at PRC-repressed and Inactive genes are 4-5 fold lower than at Active genes.

5.4. Results

5.4.1. Experimental strategy to quantitatively distinguish proteins associating with different RNAPII modification - *Universe* and *Pairwise* approach

To unravel the chromatin-bound proteome associating with different RNAPII modifications, we first set up a comprehensive experimental scheme that would robustly and specifically identify proteins and measure their enrichment relative to specific RNAPII modifications. We used the knowledge acquired in the preliminary SILAC and non-SILAC MS runs (chapters 3 and 4) to define a comprehensive and robust pChIP-SILAC experimental scheme, which included comparisons of the proteome associated with different RNAPII

modifications, the use of replicate pChIP experiments in Heavy and Light chromatin and the MS analyses of replicate forward and reverse pChIP SILAC experiments.

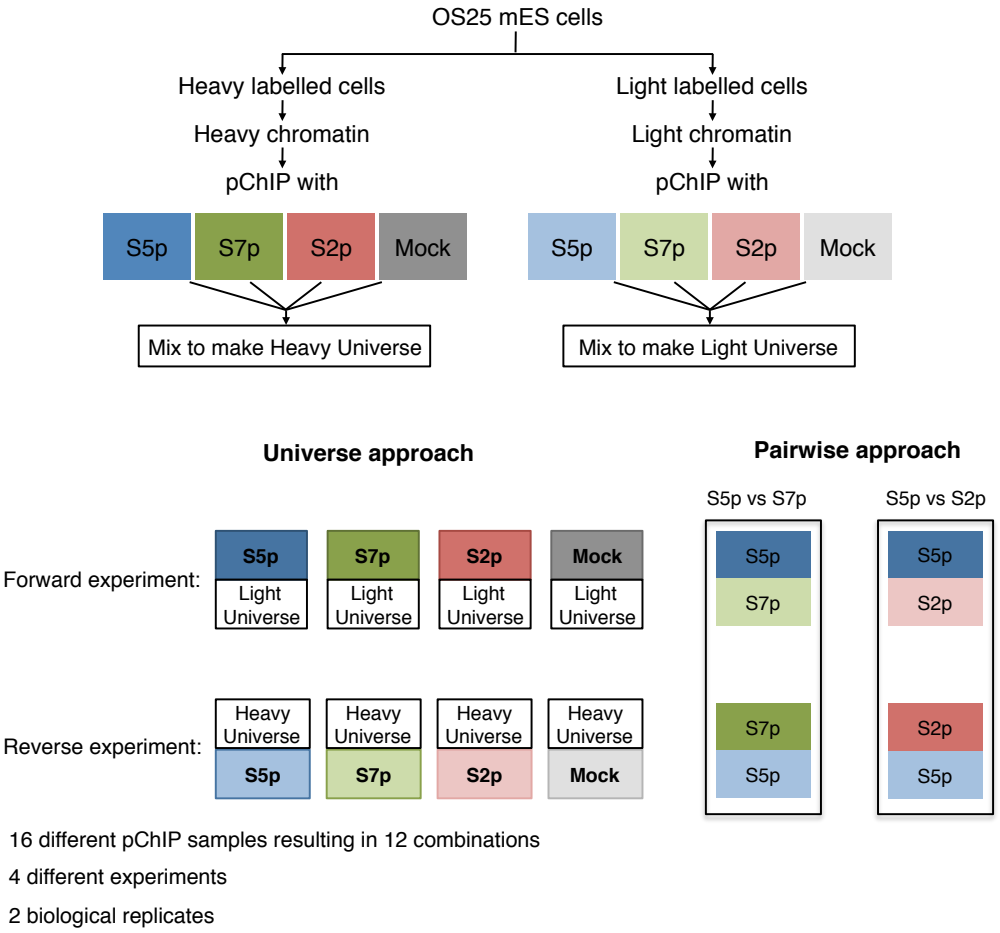
Briefly, cells were grown in SILAC conditions to make light and heavy chromatin. pChIP was performed in replicate using heavy and light chromatin and highly specific RNAPII antibodies to S5p, S7p and S2p modifications; mock IP was performed in parallel as a control. Two different complimentary approaches were devised to combine Heavy and Light pChIP samples, with different advantages in identifying proteins co-associating with chromatin occupied by RNAPII depending on its post-translational modifications.

In the first approach, termed '*Universe approach'*, we quantified the enrichment of one pChIP (heavy or light) over a defined universe sample (light or heavy, respectively). The universe sample was prepared by mixing equal volume of the four pChIP samples: S5p, S7p, S2p and mock. In the forward universe experiment, we contrasted each one of the four heavy pChIP samples with the light universe. In the reverse universe experiment, we contrasted each light pChIP with heavy universe (Fig. 5.2).

In the second approach, termed '*Pairwise approach*', we performed direct comparison between different pChIP (as in section 4.3.7) using the same pChIP eluates as used for the universe approach, and also including both forward and reverse experiments. We contrasted 'S5p with S7p' and 'S5p with S2p' in forward and reverse experiment (Fig. 5.2). S7p was not contrasted with S2p due to the complexity of the experiment, and to the fact that we were most interested in identifying proteins associated with poised states of RNAPII characterized with the presence of S5p only; both S7p and S2p are present at productively transcribed genes.

Since proteins cannot be amplified like DNA using PRC-like devices, we had to make sure that eluted pChIP material was processed extra-carefully for MS and importantly sufficient (and similar) total volume of sample was used for universe and pairwise experiments. Our experimental design was characterized by several redundant comparisons, and replicates in forward and reverse combinations of Heavy and Light samples. Performing each pChIP twice for every RNAPII modification (S5p, S7p, S2p and mock) with light and heavy chromatin resulted in a total of 16 different pChIP samples for 4 experimental schemes in 2 biological replicates (light and heavy).

The universe approach contrasts each pChIP relative to an equal mixture of four pChIP, allowing direct and robust quantification of the enrichment of peptides in a sample relative to all other samples investigated. The resulting pChIP-SILAC ratios are most stringent and enrichment is representative of the stoichiometry of protein abundance on chromatin relative to the amount of Rpb1 subunit. Owing to the highest protein complexity of these samples containing all possible proteins co-immunoprecipitated with RNAPII-bound chromatin in 4 different pChIP, we anticipated overall lower protein identification than in the pairwise approach. In the pairwise approach, only two pChIP samples are compared resulting in more robust and abundant identification of peptides and therefore proteins. Pairwise provides direct information on the two pChIP samples compared, but not the third or non-specific binders.



Universe – Four different pChIP compared

- Direct comparison between 4 pChIPs
- Relative quantification resulting in most stringent ratios
- Information on non specific binders
- Lower protein identification due to sample complexity

Pairwise – Two different pChIP compared

- Higher protein identification
- Easier and direct comparison between two modifications (no information on third)
- Lower sample complexity and higher depth of protein mapping

Figure 5.2. Comprehensive experimental setup to robustly, specifically dissect and unravel the RNAPII chromatin bound interactome. pChIP was performed on heavy and light chromatin using RNAPII antibodies (S5p, S7p and S2p) along with mock pChIP. Universe approach involved mixing each pChIP (light and heavy) into a universe sample which was contrasted with each single pChIP as described, in forward and reverse experiments. In the pairwise approach, S5p pChIP was contrasted separately with S7p pChIP and S2p pChIP, in forward and reverse experiments. Advantages and caveats in universe and pairwise approach are highlighted in the grey boxes.

5.4.2. Preliminary MS analyses of universe and pairwise pChIP mixtures for volume normalization

Before running all the universe and pairwise pChIP samples on the MS, we first analysed a small sample of the forward universe experiment by MS to determine the efficiency of Rpb1 detection and whether the Rpb1 ratios were approximately 1, or whether an adjustment of the volumes mixed from each of the 8 pChIP samples needed correction. Reassuringly, Rpb1 and Rpb2 were the proteins with the most identified peptides in all experiments highlighting the specificity of our chromatin immunoprecipitation. As previously observed in pChIP SILAC experiments (section 4.3.6), the ratios of Rpb1 were different across different experiments but consistent with Rpb2, suggesting they represent technical variability (different efficiencies of antibody precipitation, of elution or reverse crosslinking) and/or biological variability between heavy and light chromatin. Only the non-modified Rpb1 peptides identified by mass spectrometry are used for the quantification of Rpb1 proteins and its pChIP-SILAC ratio. This helps in minimising the biases associated with Rpb1 quantification (used to normalise each dataset) that potentially results from differences in CTD peptide detection due to complex post-translational modification associated with distinct chromatin states.

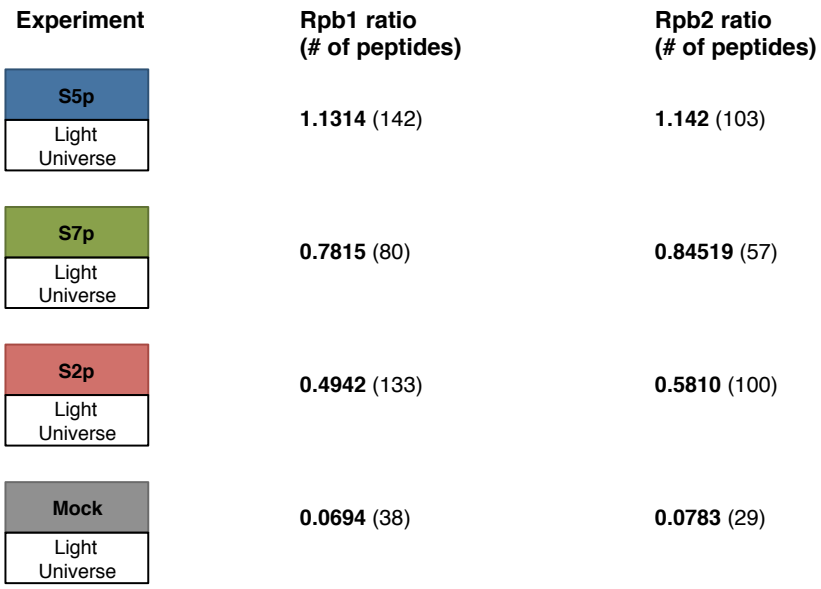


Figure 5.3. Preliminary MS analyses of the forward universe pChIP series to assess variability in pChIP-SILAC ratios. Small volumes (10 µl) of forward Universe pChIP samples were analysed in a shorter (low-depth) MS run to measure Rpb1 and Rpb2 SILAC ratio. Rpb1 and Rpb2 were robustly detected in all experiments with high peptide count.

5.4.3. Summary of pChIP-SILAC experiment run

The preliminary MS analyses of the forward universe pChIP samples showed Rpb1 and Rpb2 ratios different from 1, in particular for S2p(H)/Universe(L) (Fig. 5.3). In order to minimise the difference in pChIP-ratios, we next adjusted the volumes of individual pChIP and Universe to allow more robust ratios. Volumes were adjusted as follows:

We next set up the full pChIP experiment for high-depth MS analyses as described in the experimental setup in Fig. 5.2. Heavy and light pChIP samples were mixed together, based on their ratios in preliminary MS analyses, to make a final volume of 30 µl. Each pChIP mixture of Heavy and Light ChIP samples (12 experiments) was subsequently loaded on a 10% SDS-PAGE gel and run until 1/3rd of the gel length. After Coomassie staining 4 gel slices were cut per lane and pre-processed (de-staining, reduction and alkylation, trypsinisation, extraction of peptides and chromatographic

separation). The resulting 48 samples were analysed by MS using parameters listed in Table 5.1 and resulted in the overall identification of 737 proteins (after filtering common MS contaminants).

Table 5.1. Summary of experimental steps and parameters for MS analysis. Briefly 12 pChIP experiments were partially separated on a 10% SDS-PAGE gel and Coomassie stained, before gel fractions were cut, processed for MS analysis and analysed separately. Listed are parameters used for MS database search resulting in identification of 737 proteins.

1. Twelve pChIP SILAC samples were loaded on 10% SDS-PAGE gels and run upto 1/3 of the length of the gel.
2. Gels were coomassie stained overnight and washed to visualise protein distribution.
3. 48 gel slices (4 gel slices per experiment for 12 experiments) were cut and pre-processed for MS.

Parameters for MS analysis

MaxQuant version – 1.1.1.36
Peptide and protein FDR < 0.01
Min razor peptides – 1
Min ratio count – 1
Total MS/MS identified – 64660
Total peaks - 25248636

Total number of proteins identified – 843 proteins

Number of MS contaminants – 106 proteins

Total number of proteins for data analysis – 737 proteins

5.4.4. Steps involved in pChIP-SILAC data analysis.

For analysing the pChIP-SILAC data from our comprehensive experiment, I followed the data analysis steps summarised in Fig. 5.4. Briefly, common MS contaminants (including trypsin, keratins, heavy and light chains of antibody, and reference peptides) and proteins enriched in mock pChIP were removed. We compared two different normalizations of the pChIP SILAC ratios to analyse the dataset: to Rpb1 SILAC ratio in each MS dataset and to the average between SILAC ratios of Rpb1 and Rpb2. The former would be consistent with the fact that all antibodies used are directly targeted to modifications of the Rpb1 subunit, whereas the latter could improve the overall performance of the analysis if it made the ratio used for normalisation more robust.

Firstly, I analysed the proteins identified in the universe and pairwise approach separately to identify dependencies to specific or common RNAPII modifications. Secondly, the proteins enriched in universe and pairwise approaches were combined together in basic classification table to further delineate dependencies to RNAPII modifications. Finally, in collaboration with Prof. Mauricio Barahona and Borislav Vangelov, we have applied a systems biology approach to integrate the whole table of normalised SILAC ratios to identify 'chromatin communities' that depend on each RNAPII modification studied here (Discussed in detail in Chapter 6).

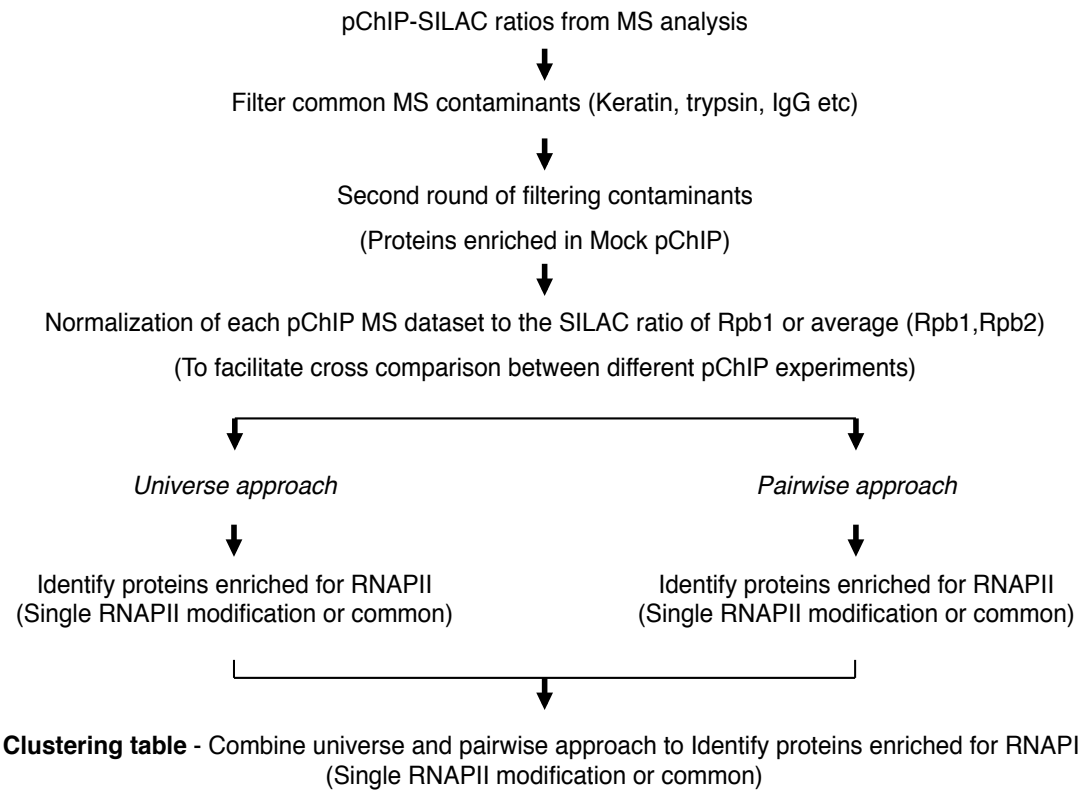
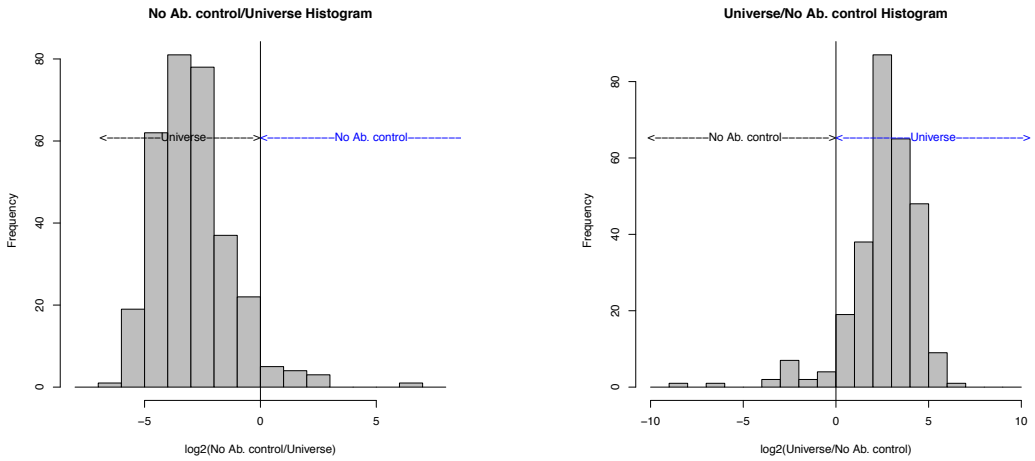


Figure 5.4. Steps involved in analysis of pChIP-SILAC data analysis to dissect dependencies to RNAPII modifications. Common MS contaminants and mock pChIP enriched proteins were filtered and pChIP ratios were normalized to Rpb1 ratios in each SILAC MS dataset. In both universe and pairwise approaches, pChIP-SILAC ratios are used to identify proteins that are enriched for a single RNAPII modification (*i.e.* S5p, S7p, S2p) or common (S5p&S7p, S5p&S2p, S7p&S2p, S5p&S7p&S2p). Finally, SILAC ratios from both approaches are combined to identify proteins dependencies relative to RNAPII modifications.

5.4.5. Filtering contaminants

To filter the proteins enriched in mock pChIP experiments, we used a stringent criteria that identifies proteins enriched in both mock pChIP experiment (pChIP experiment contaminants). Histograms of pChIP-SILAC ratios measured in the mock pChIP datasets highlight that most proteins are specifically enriched for the 'Universe', whereas very few proteins are enriched in the mock pChIP, in both forward and reverse mock-Universe pChIP-SILAC experiments (Fig. 5.5). To remove mock-pChIP contaminants, a stringent ratio cut-off of 1.0±0.2 (forward ratio $\geq$ 0.8 and reverse ratio $\leq$ 1.2) identified 32 proteins that were removed from all datasets (including the pairwise comparisons). These results confirm the specificity of our ChIP protocol and RNAPII antibodies in immunoprecipitating chromatin bound by RNAPII (see also DNA-ChIP results, Figs. 3.6 and 4.4). Several proteins removed with this stringent threshold appeared irrelevant (such as desmoplakin and junction plakoglobin), but other proteins appeared interesting, which could have accidentally been identified in the mock pChIP with a larger SILAC ratio due to their respective lower number of peptides (peptide h/l count). The latter group included five interesting proteins: Mybbp1 (essential for mES cell pluripotency; Universe/mock ratio = 0.002; peptide h/l count =1), Baf190 (chromatin remodeller; Mock/universe ratio = 1.946; peptide h/l count =1), Pcna (proliferating cell nuclear antigen; Mock/universe ratio = 6.6138; peptide h/l count =2) and RNA helicases (Dhx38 and Ddx1; Mock/universe ratio = 1.2172 and 6.605; peptide h/l count =1 for both). This filtering step was nevertheless kept to minimise biases in the pChIP dataset analyses. As with any MS analysis, lower abundancy proteins in particular with lower number of detectable peptides (e.g. if they are small) are more difficult to detect and therefore can be missed.



Non specific binders	Criteria	Number of proteins
Forward ratio (Mock/Universe)	≤ 0.8	13
Reverse ratio (Universe/mock)	≥ 1.2	17
Protein with ratio only in mock pChIP		2
Total number of proteins		**32**

Figure 5.5. Identification of contaminant proteins enriched in mock pChIP relative to Universe. Stringent criteria were applied to identify proteins enriched in mock-pChIP based on SILAC ratios from Forward and Reverse mock-Universe experiments. Forward (≤ 0.8) and reverse (≥ 1.2) threshold ratios were applied to identify proteins enriched in the mock-pChIP datasets relative to the universe mixture, identifying 32 proteins which were filtered out from all MS datasets (universe and pairwise).

5.4.6. RPB1 normalization

Before starting to analyse the pChIP-SILAC dataset, we first explored the options to normalise the data given the comprehensive, complimentary experimental setup and for effective comparison of both universe and pairwise approach.

It is important to note that we have used highly specific antibodies against Rpb1-CTD of RNAPII and a ChIP protocol carefully optimised to produce good chromatin yield with minimum background (Xie and Pombo 2006; Stock *et al.* 2007b; Brookes *et al.* 2012). Given that all modifications occur on Rpb1-CTD and that our aim is to understand association of proteins relative to RNAPII, it

seemed most intuitive to normalise all SILAC ratios to the SILAC ratio of the Rpb1 subunit of RNAPII in each MS dataset. However, protein detection by MS depends on its amino-acid composition and length, as it depends on proteolysis of lysine peptide bonds and the size of the resulting peptides. To determine whether Rpb1 is robustly detected in our MS analyses, we measured the number of Rpb1 peptides used to calculate the Rpb1-SILAC ratio in each dataset (Fig. 5.6; all Rpb1 peptides were used including any detected unmodified CTD peptides). The lowest number of Rpb1 peptides (34 peptides) was detected in "S7p/Universe" and this experiment had overall the lowest number of proteins identified (230 proteins in comparison with ~550 proteins in all the other datasets). Looking at the distribution of Rpb1 peptides, most have SILAC ratios close to the log of median Rpb1 ratio, and few peptides were specifically enriched in either heavy or light, resulting in outlier ratios (Fig. 5.7). Some CTD peptides were distinctly enriched either in heavy or light samples (e.g. 'YSPTSPTYSPTSPK', 'YSPTSPTYSPTSPVYTPTSPK') indicating the possibility of post-translational modification on these peptides. Although analysis of protein modification is an obvious exciting further development of pChIP, we have not at this stage expanded our analyses of MS spectra to explore this issue; our superficial analyses indicate this will certainly be possible, but due to the proteome complexity observed in each pChIP dataset, may require more targeted MS analyses of specific proteins.

Experiment	Peptide count (h/l) Forward experiment (CTD peptides)	Reverse experiment (CTD peptides)
S5p Universe	180 (6)	194 (8)
S7p Universe	34 (1)	195 (6)
S2p Universe	74 (4)	129 (2)
S5p S7p	113 (10)	104 (4)
S5p S2p	106 (5)	115 (3)

Figure 5.6. Robust detection of Rpb1 peptides across all MS SILAC datasets. Rpb1 (polr2a) subunit of RNAPII was robustly detected in all pChIP experiments (forward and reverse) with a large number of peptides detected in both heavy and light samples (represented by peptide count). All detected unmodified Rpb1 peptides were used for calculating pChIP ratios. The unmodified CTD peptides detected in each experiment are also listed.

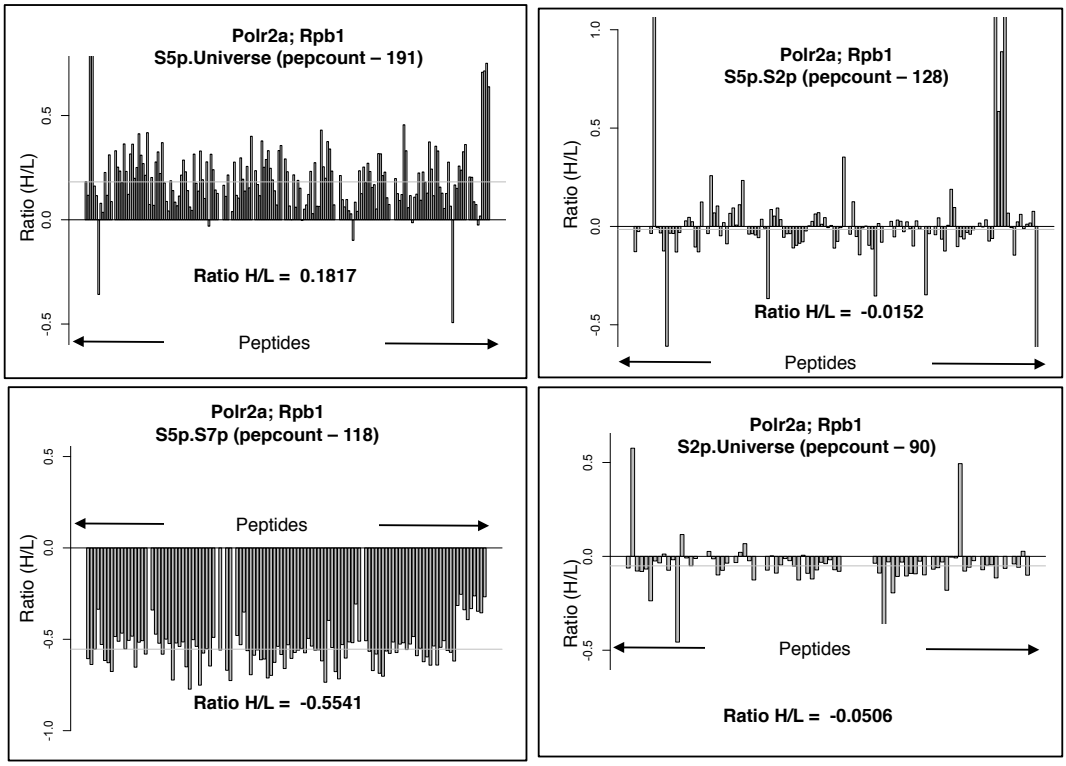


Figure 5.7. Detailed analysis of pChIP SILAC ratios for Rpb1 and its peptides. Peptides corresponding to Rpb1 subunit were plotted in order (from N- to C-terminal) as barplots to observe the distribution of their H/L enrichment in specific pChIP H/L mixtures. Most peptides follow a similar distribution within each experiment with only a few outliers deviating from the log of the median of SILAC ratio used for the whole subunit. Grey line represented the MS reported pChIP ratio for Rpb1 (Also displayed as Ratio H/L). Numbers of identified peptides are shown in the title of each graph (pepcount) along with experiment name (as H/L). X-axis represents the different Rpb1 peptides detected in both heavy and light pChIP in order from N- to C-terminus, Y-axis represents peptide ratios in log scale.

I next asked whether other Rpb subunits were also well detected after the pChIP experiments with antibodies against Rpb1 modifications, and how their SILAC ratios were distributed relative to Rpb1 ratio. In all SILAC MS datasets, 4-8 RNAPII subunits were detected, and their ratios were relatively close to the Rpb1 ratio. For example, plotting the distribution of all protein ratios in a scatter plot for three of the Pairwise MS datasets (Fig. 5.8) shows that most of the detected RNAPII subunits have SILAC ratios close to Rpb1 (axis at 0,0). Rpb subunits with fewer peptides, such as Rpb7 (Polr2g) have SILAC ratios more distant from the Rpb1 ratio, as expected (not shown).

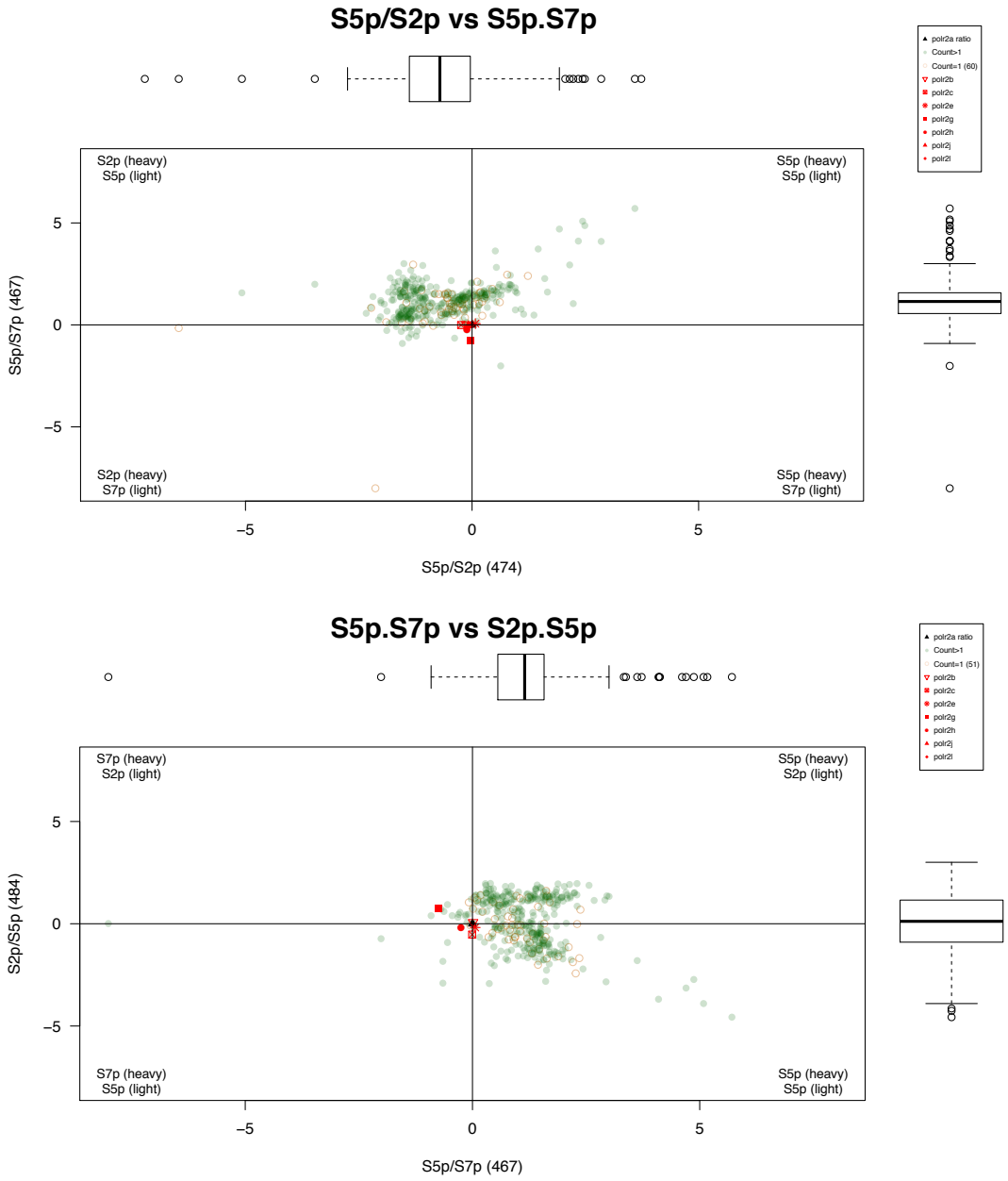


Figure 5.8. Position of pChIP SILAC enrichments from detected Rpb subunits relative to Rpb1. Scatter plot for proteins identified in pairwise approach. Rpb1 subunit is located at (0,0) and marked by black triangle. Other Rpb subunits lie mostly adjacent to Rpb1. Number of proteins identified and their distribution is shown in axis labels and as box plots (next to axes)

5.4.7. Proteome of RNAPII-bound chromatin; universe approach

To begin dissecting the protein dependencies to RNAPII modifications, I first started to analyse the universe dataset, based on whether proteins showed ratios H/L >1 in the light universe datasets or ratios H/L <1 in the heavy universe datasets. In these initial analyses of the pChIP datasets, proteins that consistently appeared enriched relative to the pChIP experiments of only a single RNAPII modification and not the other two modifications, across all 6 Universe datasets, were considered 'specific' for association with chromatin enriched for RNAPII containing this modification. The numbers of proteins enriched for S5p, S7p and S2p, or shared between modifications are represented as a Venn plot in Fig. 5.9.

As expected from the observation that all three studied Rpb1 modifications (S5p, S7p and S2p) are simultaneously present at RNAPII transcribing active genes (Brookes et al 2012), we observed that the majority of proteins are shared between S5p and S2p (155 proteins), or between S5p, S7p and S2p (24 proteins). We also find a large cohort of proteins enriched only in pChIP for S2p (197 proteins). Very few proteins are detected enriched in the S7p datasets relative to Universe (18 proteins), but interestingly this small group appears biologically relevant as it includes five transcription initiation factor subunits (TAFs) and the transcription pausing factor (Nelf-D, negative regulator of transcription elongation), all proteins with known roles in the transition between initiation and productive elongation of RNAPII, a role also recently associated with the S7p modification. Interestingly, I detected 183 proteins enriched only relative to S5p in the S5p/Universe datasets and 197 proteins in the S2p/Universe datasets. The former could consist of proteins specifically associated with RNAPII-S5p^{+}S7p^{-}S2p^{-} at Polycomb repressed genes (Brookes *et al.* 2012), and the latter with proteins associated with elongating RNAPII, which although marked by S5p, may be more effectively immunoprecipitated with the S2p antibody.

To explore the lists of proteins in each group, I performed GO analyses (Fig. 5.9). Proteins common to all modifications (*i.e.* co-associated with RNAPII-bound chromatin independently of S5p, S7p or S2p) were enriched for histones (GO term: 'nucleosomal assembly'), proteins in the S7p&S2p, S5p&S2p and S2p groups were characterized by GO terms related with active transcription and co-transcriptional RNA processing such as 'mRNA processing', 'RNA processing', 'RNA transport', 'transcription' and 'H2B monoubiquitination', consistent with the fact that S2p is a mark of transcriptional elongation. Ribosomal proteins were also identified in this group (GO term 'translation', consistent with reports for nuclear translation (Iborra *et al.* 2004; David *et al.* 2012), and biochemical purification of ribosomes with RNAPII (Das *et al.* 2007; Moller *et al.* 2012). Remarkably, proteins in the S5p-only group were enriched for a more diverse range of processes. Most significant terms included 'Negative regulation of transcription', 'DNA replication' and 'chromatin remodelling' (Fig. 5.9).

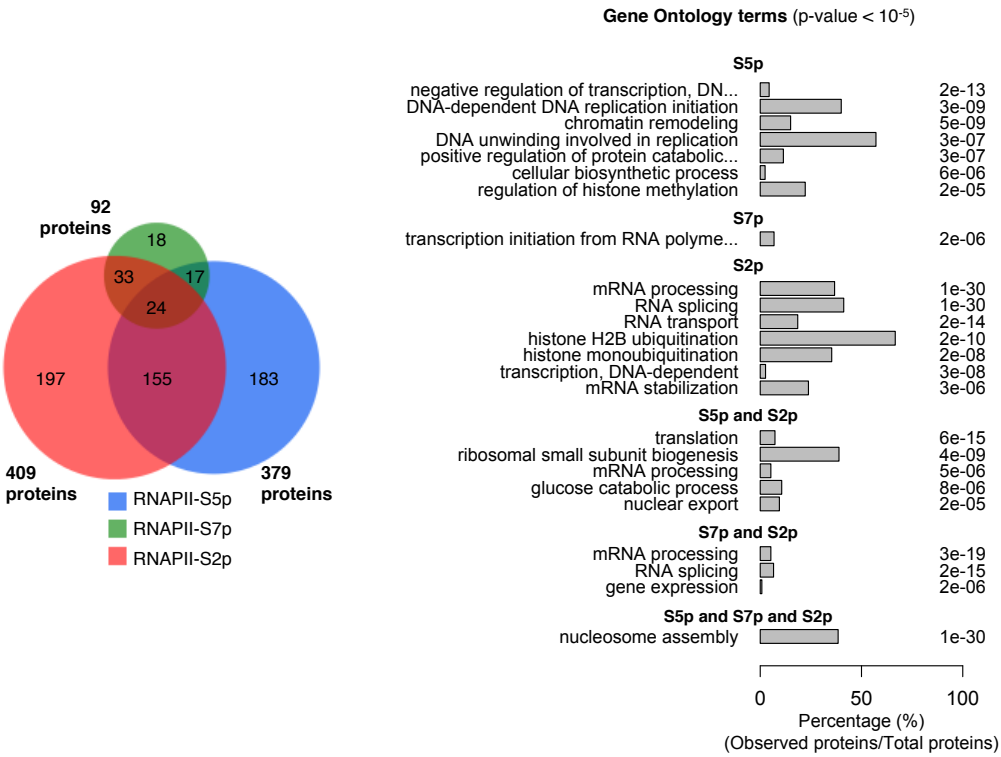


Figure 5.9. Classification of the chromatin-bound proteome that co-exists with different RNAPII modifications from the universe approach datasets. Left, Venn diagram representing proteins with SILAC ratios consistently enriched for a specific RNAPII modification, or shared between modifications. Right, significant GO terms (p-value $<10^{-6}$) for each group of proteins enriched in different category; no GO terms were enriched in S7p-only due to the low number of proteins in this group.

5.4.8. Proteome of RNAPII-bound chromatin; pairwise approach

I applied similar data analysis steps to ask which proteins were enriched only for S5p, S7p, S2p and those shared between RNAPII modifications in the 4 pairwise approach MS datasets (S5p.S7p and S5p.S2p in forward and in reverse h/l combination). In the S5p/S7p datasets (Fig. 5.10), most proteins are enriched for S5p (415 proteins), and very few for S7p only (28 proteins); 57 proteins had SILAC ratios, between 0.8 and 1.2, therefore not consistently enriched relative to a single modification and were therefore considered as common to the two modifications. The total number of S7p proteins (common and S7p only) was 85 and it was very similar to the number of S7p proteins identified from universe approach (92 proteins; Fig. 5.9). This raises an important question whether the low numbers of proteins detected after S7p

immunoprecipitation are due to technical or biological constraints. The former could be due to lower efficiency of the immunoprecipitation with the S7p antibody, whereas the latter could result from S7p existing in a very short window of the transcription cycle, and co-associating with a smaller number of proteins. The latter would be consistent with the perceived role of S7p in the transition between initiation and productive elongation (Czudnochowski *et al.* 2012).

As observed previously for the universe approach, GO analysis showed that common proteins were enriched for translation and mRNA processing terms, whereas S5p only proteins were involved in mRNA processing, RNA splicing, RNA export and DNA replication (Fig. 5.9). The detection of GO terms related to productive elongation, is likely due to the fact that elongating RNAPII is marked by S5p, and in this pairwise comparison of S5p with S7p, we do not distinguish proteins more highly associated with S5p-only or S2p-only. In relation to S7p-only proteins, due to the low number of proteins detected (28), GO analysis did not yield specific terms, but interestingly this small group of proteins again included Transcription initiation factor subunits (TAFs), Transcription pausing factor (Nelf-D) and CDK-activating kinase assembly factor (Mnat1; essential for RNAPII promoter escape and elongation), proteins known to act at the transition between RNAPII initiation and elongation, some of which are also detected as S7p specific in the universe pChIP approach.

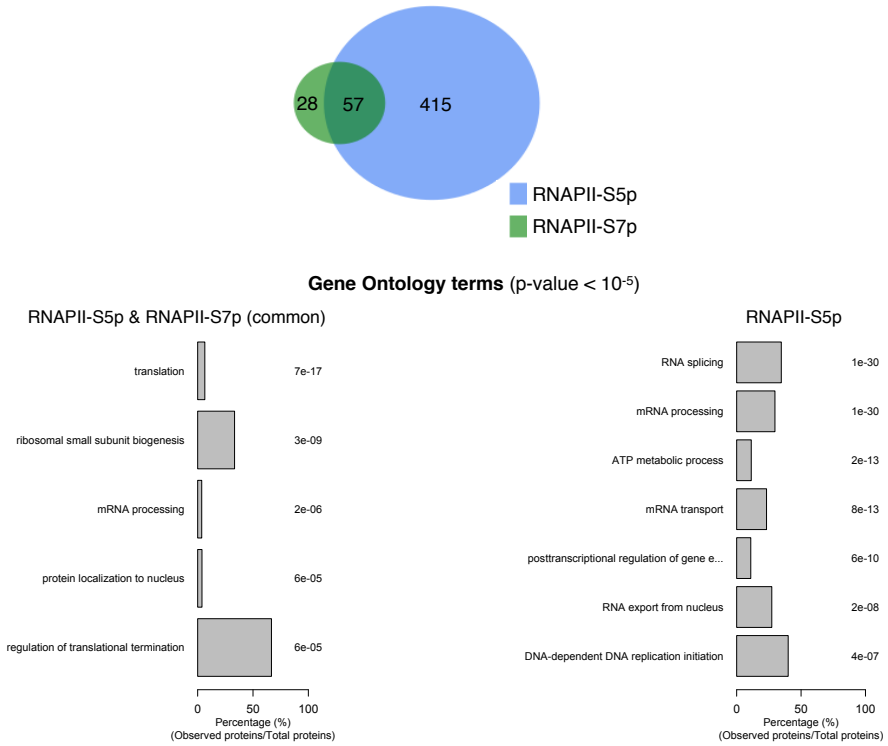


Figure 5.10. Identifying proteins dependencies to RNAPII-S5p and/or RNAPII-S7p from pairwise experiments. Proteins were classified either enriched for S5p, S7p or common based on pChIP-SILAC ratios and represented as a Venn diagram. Significant GO terms (p-value $<10^{-5}$) for proteins enriched in S5p-only and common S5p.S7p. No significant terms were observed for proteins enriched in S7p-only.

Performing analysis on S5p/S2p pairwise approach datasets identified almost 30% more proteins than in the S5p/S7p datasets (686 compared to 500 proteins). We identify 249 proteins not specifically enriched in S5p or S2p (*i.e.*, shared between the two marks), 287 proteins enriched for S2p and 150 for S5p. As with the S5p.S7p pairwise experiment, this analysis is blind to how specific proteins associate with the S7p mark.

GO analysis (Fig. 5.11) showed that the group of proteins enriched for S2p or common to S2p and S5p is characterised by terms related with productive mRNA transcription, *i.e.* 'mRNA processing', 'RNA splicing' and 'translation', as seen previously, while the group of proteins enriched for S5p-only also resulted in more diverse GO terms, such as 'DNA replication', 'chromatin modifications', and 'chromatin remodelling'.

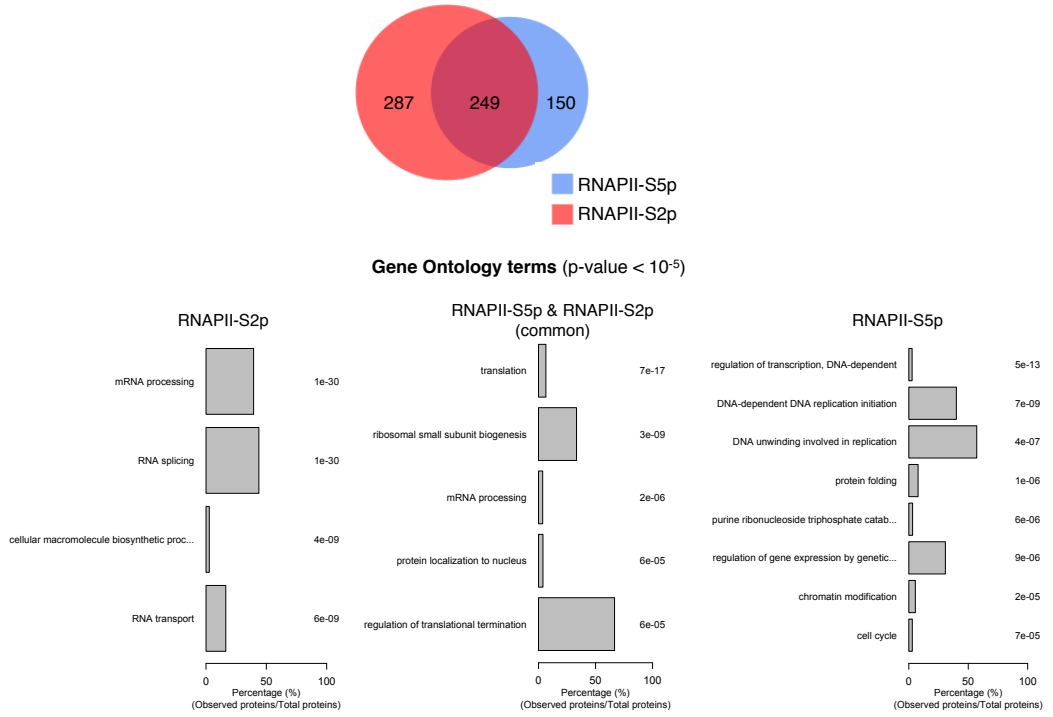


Figure 5.11. Identifying proteins dependencies to RNAPII-S5p and/or RNAPII-S2p from pairwise experiment. Top, Proteins were classified either as enriched for S5p, S2p or shared, based on pChIP-SILAC ratios, and represented as a Venn diagram. Bottom, Significant GO terms (p-value <10^{-5}) for proteins enriched in S5p-only, S2p-only and common.

5.4.9. Combining universe and pairwise approaches

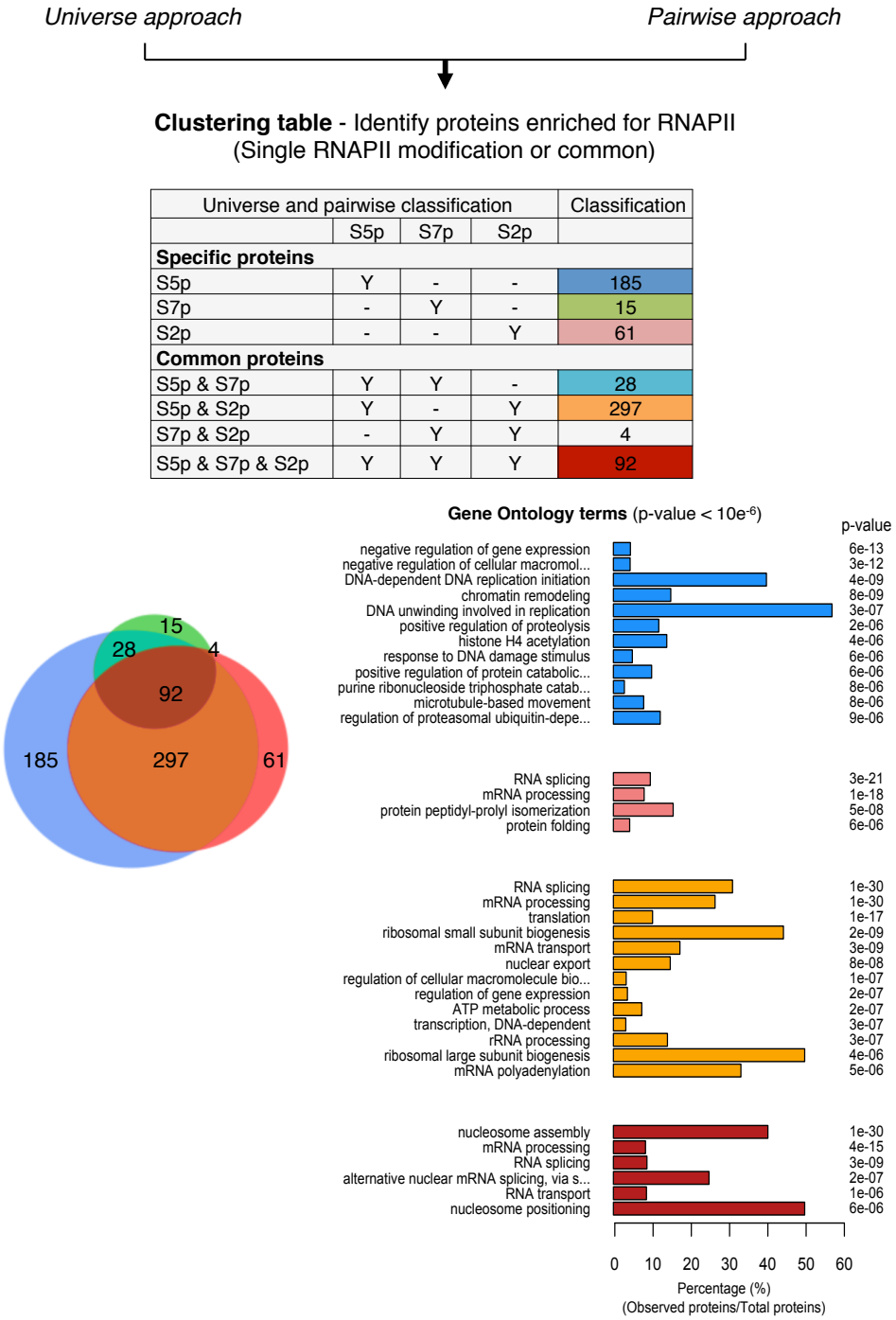
From the overview of the universe and pairwise approach results, both approaches gave similar results in terms of number of proteins and significant GO terms. We next asked how would the distribution of proteins and their dependencies change if we combined both the analyses.

Taking all the proteins represented in universe and pairwise Venn diagrams, we classified proteins according to a simple classification table (Fig. 5.12, top) which identifies proteins consistently enriched for only one RNAPII modification across all replicate datasets, shared between two RNAPII modifications or shared between all three modifications (Fig. 5.12, bottom left). We observe 92 proteins shared between all RNAPII modifications, 297 proteins shared between S5p and S2p, 28 proteins shared between S5p and

S7p, 61 proteins specifically associating with S2p and 185 proteins specifically associating with S5p.

Looking at significant GO terms, we observe that all groups of proteins related with S2p, either alone or with the other marks (S5p&S7p&S2p, S5p&S2p and S2p) are characterized with 'mRNA processing', 'RNA splicing', 'nucleosome assembly' and 'RNA transport', processes downstream of transcription initiation as expected for the S2p mark. Interestingly, 'translation' and 'ribosome biogenesis' are two terms that are consistently enriched with S5p&S2p shared proteins, and may be related with observations of transcription-dependent translation in association with RNAPII transcription sites (Iborra *et al.* 2001).

As hinted in the previous separate analyses of universe and pairwise approaches, the group of S5p-only proteins have GO terms markedly different from those seen in association with S2p, and productive RNAPII elongation. The most significant term is 'negative regulation of gene expression', which includes Polycomb proteins, previously shown to co-occupy chromatin bound by RNAPII-S5p^{+}S2p^{-}, by sequential DNA-ChIP experiments for a small number of genomic regions (Brookes *et al.* 2012). In the S5p-only group of proteins, we also consistently enrich for 'chromatin remodellers', and perhaps more unexpectedly for 'DNA replication' and 'metabolic terms'.



Universe and pairwise classification				Classification
	S5p	S7p	S2p	
Specific proteins				
S5p	Y	-	-	185
S7p	-	Y	-	15
S2p	-	-	Y	61
Common proteins				
S5p & S7p	Y	Y	-	28
S5p & S2p	Y	-	Y	297
S7p & S2p	-	Y	Y	4
S5p & S7p & S2p	Y	Y	Y	92

Figure 5.12. Combining universe and pairwise approaches to dissect proteins dependencies to RNAPII modifications. Proteins identified and enriched from universe and pairwise approach were combined together in a simple classification table. Protein enriched only in S5p from both universe and pairwise approach datasets was classified as S5p only and similarly for S7p, S2p and common proteins. Venn diagram represents classification of proteins with respect to RNAPII modifications and significant GO terms for the enriched proteins.

5.4.10. Examples of proteins associating with RNAPII as identified by pChIP.

We next analysed in more detail some of the important GO terms and the distribution of SILAC ratios for specific proteins across our comprehensive pChIP-SILAC experiment. We first looked at distribution of exemplars of proteins that were detected in our experiment, interact with RNAPII and were also identified in all the ten separate SILAC MS datasets. We also discuss proteins with varying pChIP ratios across pChIP experiments that are not enriched for any RNAPII modification.

5.4.11. Chromatin remodellers

During the process of transcription, RNAPII is recruited to promoters of active genes having open chromatin architecture characterised by tri-methylation of lysine residue at position 4 of histone H3 (H3K4me3). H3K4me3 is also present at PRC-repressed genes along with repressive H3K27me3 and H2Aub1 catalysed by Polycomb proteins.

In our pChIP datasets, we observe selected chromatin remodellers selectively associating with RNAPII-S5p. These subunits are part of multi-protein complexes that include NuRD, Swi/Snf (BAF and PBAF) and CHRAC chromatin-remodelling complex. Plotting the distribution of pChIP-SILAC ratios demonstrates consistent enrichment of S5p from universe and pairwise experiments while no enrichment for S7p or S2p (Fig. 5.13). Other chromatin remodellers that associate with RNAPII-S5p include Rbbp4/7, Mta2, Arid1a and Baz1b.

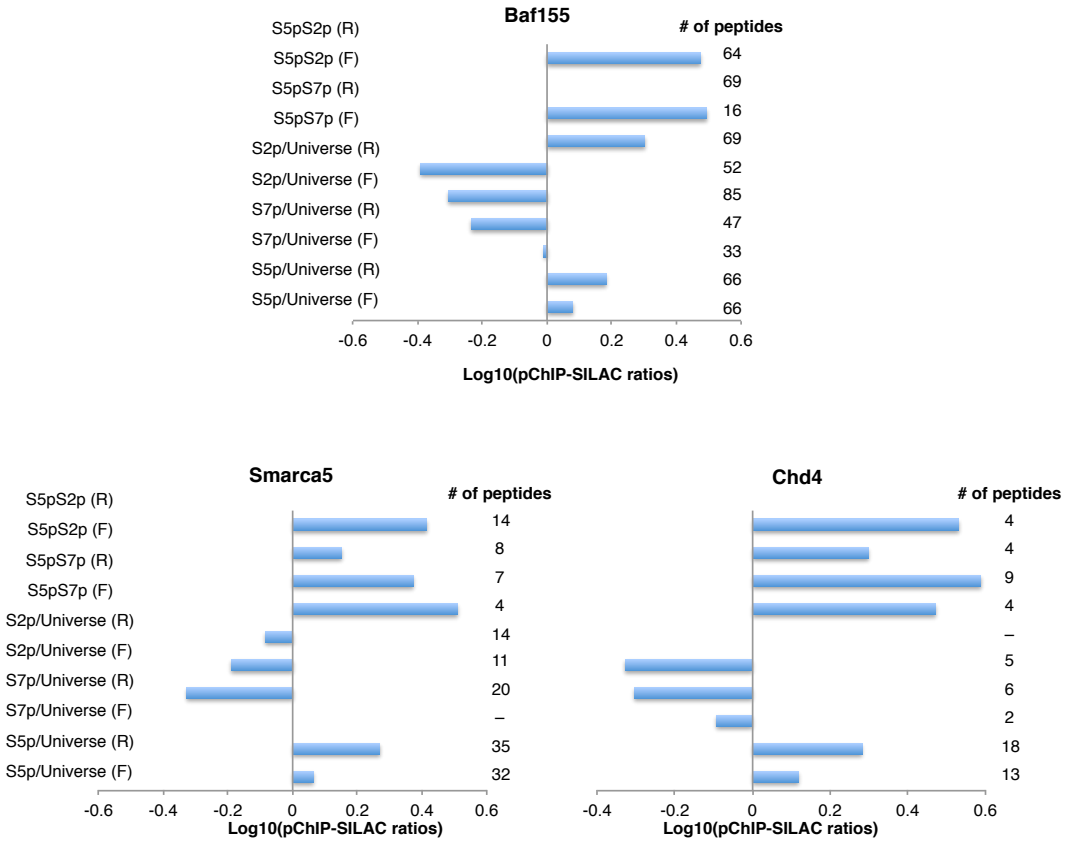


Figure 5.13. Chromatin remodellers also specifically co-exist on chromatin with RNAPII-S5p. Barplot for some chromatin remodellers (Baf155, Smarca5 and Chd4) that specifically co-exist with RNAPII-S5p. Proteins with positive log ratios are enriched heavy pChIP while proteins with negative log ratios are enriched for light pChIP. pChIP-SILAC ratios are plotted in log scale and number of peptides are represented next to ratios on barplot.

5.4.12. Polycomb proteins

Our lab has previously identified RNAPII-S5p^{+}S7p^{-}S2p^{-} selectively co-existing with Polycomb proteins on chromatin after sequential-ChIP across ~20 genomic regions analysed by qPCR (Stock *et al.* 2007b; Brookes *et al.* 2012). Across our pChIP datasets, Polycomb proteins are not as abundantly detected as the MCM proteins discussed above, but when detected they selectively associate with RNAPII-S5p only, and not S2p or S7p. These proteins are in general detected with lower peptide count, probably due to their smaller size or potentially due to incomplete separation in chromatography or MS charge detection on proteins. We identify 4 Polycomb

proteins (Ring1b, Eed, Suz12 and Jarid2; with peptide h/l count of 3, 17, 5 and 10, respectively) in our datasets and, even at low peptide detection, we see clear association only with RNAPII-S5p and not with S7p or S2p.

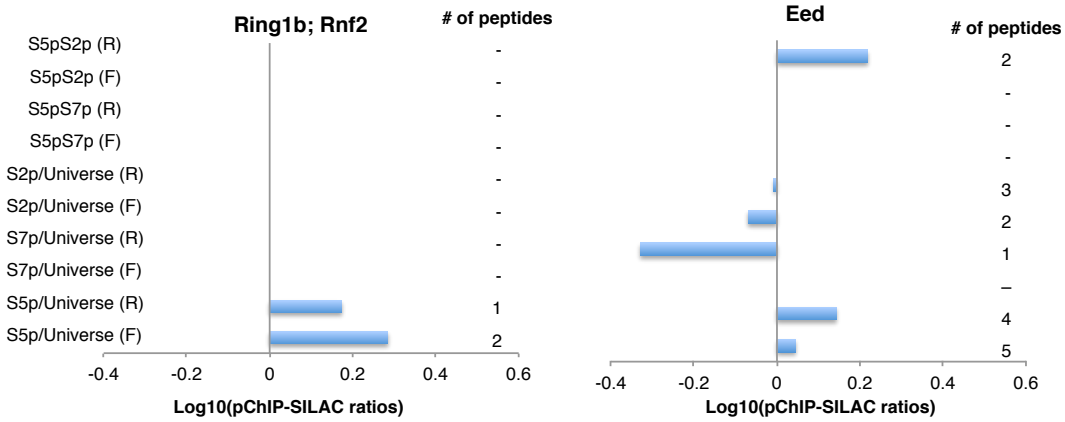


Figure 5.14. Polycomb proteins associate on chromatin with RNAPII-S5p. Barplot for Polycomb proteins (Ring1b and Eed) and their pChIP-SILAC ratios highlight specific association on chromatin with RNAPII-S5p, albeit at low peptide count. Other core Polycomb proteins (Jarid2 and Suz12) were also identified as RNAPII-S5p only; note consistent enrichment detected in pair-wise experiments characterized by higher ability to produce measurable H/L ratios for less abundant proteins associated with a specific modification. H/L SILAC ratios were inverted in the reverse (to be similar to forward) experiments and represented as indicated. pChIP-SILAC ratios are plotted in log scale and number of peptides are represented next to ratios on barplot.

5.4.13. DNA replication

Stem cells proliferate rapidly and actively replicate to produce cells that are pluripotent and undifferentiated. Understanding replication in stem cells has been a major focus in the field and several papers have put forward ideas linking transcription and replication in mammalian cells (Jackson and Pombo 1998; Hiratani *et al.* 2008). We identify all components of the Mcm2-7 complex selectively associating on chromatin along with S5p. Looking at the distribution of pChIP-SILAC ratios across different experiments, we observed that Mcm proteins were enriched for S5p in universe and pairwise experiments while being depleted for S7p and S2p consistently in pairwise and universe approaches (Fig. 5.15).

RNAPII-S5p (without S7p and S2p) is found at approximately 25% of the RefSeq gene promoters, and many of their coding regions, in mES cells (Brookes *et al.* 2012). For most of their short cell cycle, with average doubling time ~12 hours, ES cells are in S phase, having a very short G1 phase (Udy *et al.* 1997; Azuara *et al.* 2006). Given this, the simultaneous association of RNAPII-S5p and replication proteins on chromatin raises an interesting prospect of what comes first and how it influences mES cell architecture and dynamic cellular organisation.

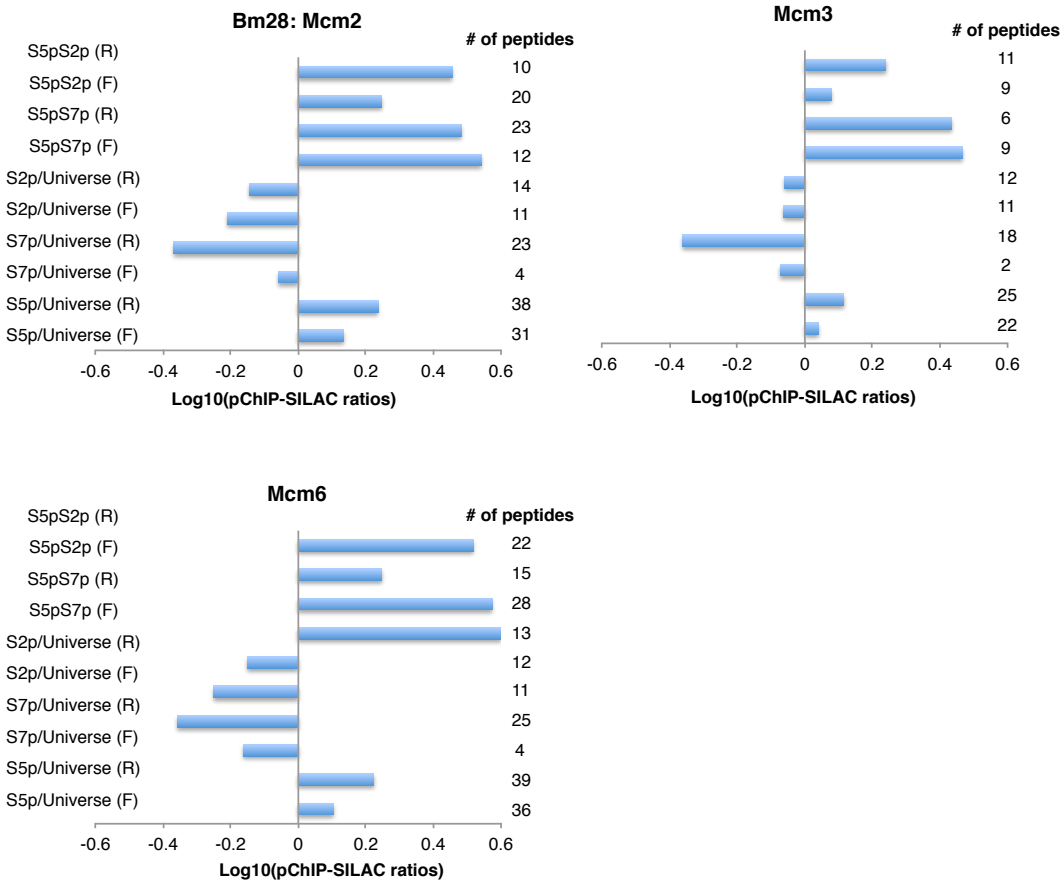


Figure 5.15. MCM2-7 complex is robustly identified across pChIP datasets and is enriched specifically on chromatin containing RNAPII-S5p. Barplot for three MCM proteins and their pChIP-SILAC ratios highlight specific association on chromatin with RNAPII-S5p. H/L SILAC ratios were inverted in the reverse (to be similar to forward) experiments and represented as indicated. pChIP-SILAC ratios are plotted in log scale and number of peptides are represented next to ratios on barplot.

5.4.14. Ribosomal proteins

Ribosomal proteins are very abundant and are often considered contaminants and filtered during MS analysis of immunoprecipitates (e.g. (Das *et al.* 2007). In our cohort of pChIP datasets obtained after in vivo crosslinking of chromatin, we consistently detect ribosomal proteins, found more significantly associated with chromatin co-occupied with RNAPII-S5p and S2p. RNAPII-S5p and -S2p are hallmarks of active transcription (initiation and elongation), with S5p being present at both promoters and coding regions of active genes, whereas S2p is abundant through coding regions. Ribosomes (and translation) have previously been described in association with active transcription sites in mammalian cells (Iborra *et al.* 2001), and in the interbands of Drosophila polytene chromosomes known to contain active genes (Brogna *et al.* 2002). The specific enrichment of ribosomal and translation GO terms with S5p and S2p suggests that we are capturing similar phenomenon of nuclear transcription-coupled-translation in mES cells. The concept of nuclear translation has been challenged and debated (Dahlberg *et al.* 2003), but reinforced in a recent study (David *et al.* 2012).

5.4.15. S2p-associated proteins

In our SILAC MS dataset, we also observe some proteins specifically enriched for S2p only. Interestingly, all these proteins are involved in splicing, mRNA processing and RNA processes and astonishingly these proteins have no peptides detected in the S5p/S7p pairwise experiment (both forward and reverse), thereby reinforcing evidence for a specific association with S2p.

Prpf4 and Prpf38b are pre-mRNA splicing factors, central components of spliceosome (Uniprot 2012). Prpf4 is part of U5 snRNP complex and also interacts with splicing dependent exon-junction complexes (EJC). Prpf4 and Prpf38b have identical patterns of pChIP ratios and were enriched for S2p in both the universe dataset (no enrichment for S5p or S7p universe datasets) and enriched for S2p in pairwise dataset (S5p.S2p).

Cpsf3 is a component of cleavage and polyadenylation specificity factor (CPSF) complex recruited at the 3'UTRs of pre-mRNA and fundamental to 3'end formation. Cpsf3 was enriched for S2p in universe dataset (S2p.Universe and no enrichment for S5p or S7p) as well in pairwise dataset (S5p.S2p). Cbp80 is a component of the cap-binding complex (CBC), which co-transcriptionally binds to pre-mRNAs (5' cap; (Izaurralde *et al.* 1994). It is involved in various processes such as pre-mRNA splicing, translation regulation, nonsense-mediated mRNA decay, RNA-mediated gene silencing (RNAi) by microRNAs (miRNAs) and mRNA export. Consistent with other S2p proteins, pattern for Cbp80 was similar and enriched for S2p in universe and pairwise dataset.

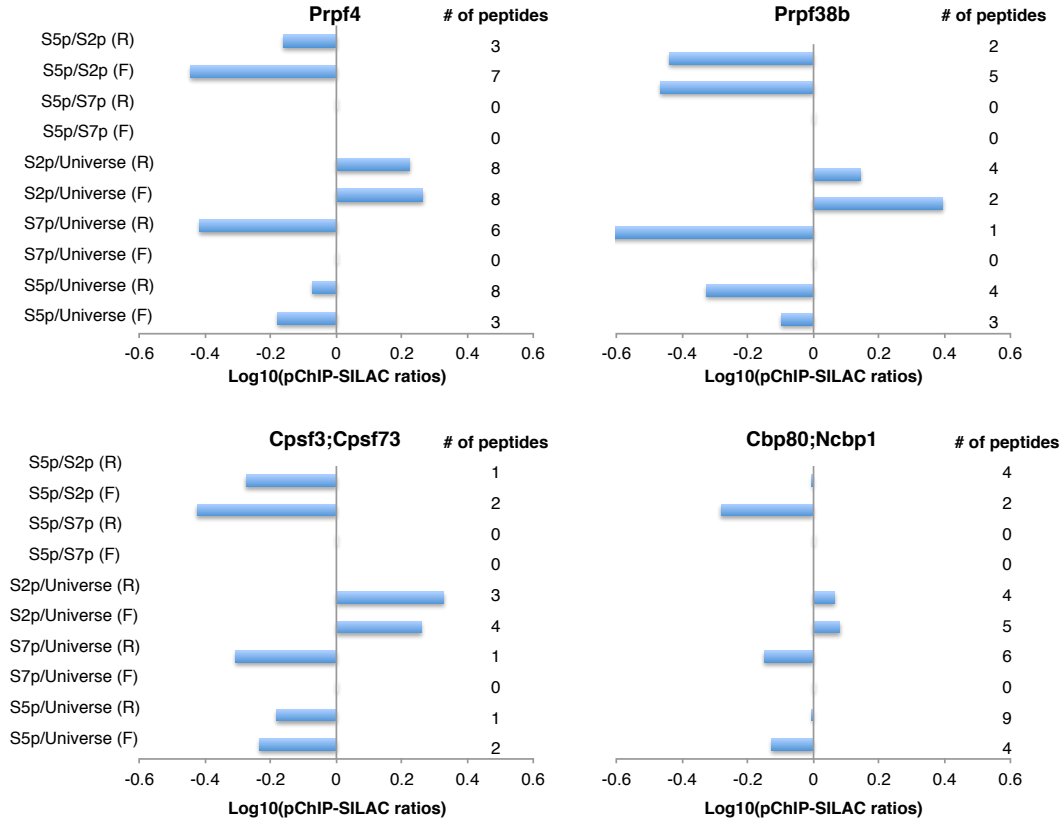


Figure 5.16. Examples of proteins associated with RNAPII-S2p on chromatin. Barplot for S2p associating proteins and their pChIP-SILAC ratios across 10 pChIP MS datasets (clockwise from left top; Prpf38b, Cbp80, Cpsf3 and Prpf4). Proteins with positive log ratios are enriched heavy pChIP while proteins with negative log

ratios are enriched for light pChIP. pChIP-SILAC ratios are plotted in log scale and number of peptides are represented next to ratios on barplot.

5.4.16. Proteins without any specific enrichment for RNAPII modifications

In our pChIP MS dataset, we also observed some proteins that had no specific enrichment for RNAPII modifications. These included Paf1 complex among other proteins. Paf1 complex includes Paf1, Cdc73, Leo1, Ctr9, Rtf1 and Wdr61 proteins and has multiple functions during RNAPII transcription and is also fundamental to mES cell pluripotency (Ding *et al.* 2009; Uniprot 2012). The Paf1 complex interacts with both un-phosphorylated RNAPII-CTD and S5p/S2p RNAPII and is involved in transcriptional elongation, histone H2B, mRNA 3' end formation and polyadenylation (Ding *et al.* 2009). We observed Paf1 complex subunits having varying pChIP-SILAC ratios within the forward and reverse experiments and in addition without much enrichment across 10 MS datasets (Fig. 5.17; Log10 ratios within ± 0.2). As expected, RNAPII subunits also exhibit similar behaviour i.e. no enrichment for specific RNAPII modifications and in addition are not particularly enriched across 10 MS datasets.

The Paf1 complex (and other proteins with similar behaviour) has a pattern of pChIP-SILAC ratios across 10 MS experiments that is closely similar to the ratios of the RNAPII subunits, being therefore more closely affected by Rpb1 normalization (see also Fig. 5.8). Depending on our normalization, these proteins accordingly affected and therefore exhibit a variation in pChIP-SILAC ratios that accompanies the RNAPII subunits. Cases such as this one justify our subsequent use of a systems biology approach in which proteins are clustered based on their SILAC enrichments and not simply on the direction of enrichment, a procedure that by definition can more successfully measure coincidences in protein association and stoichiometry

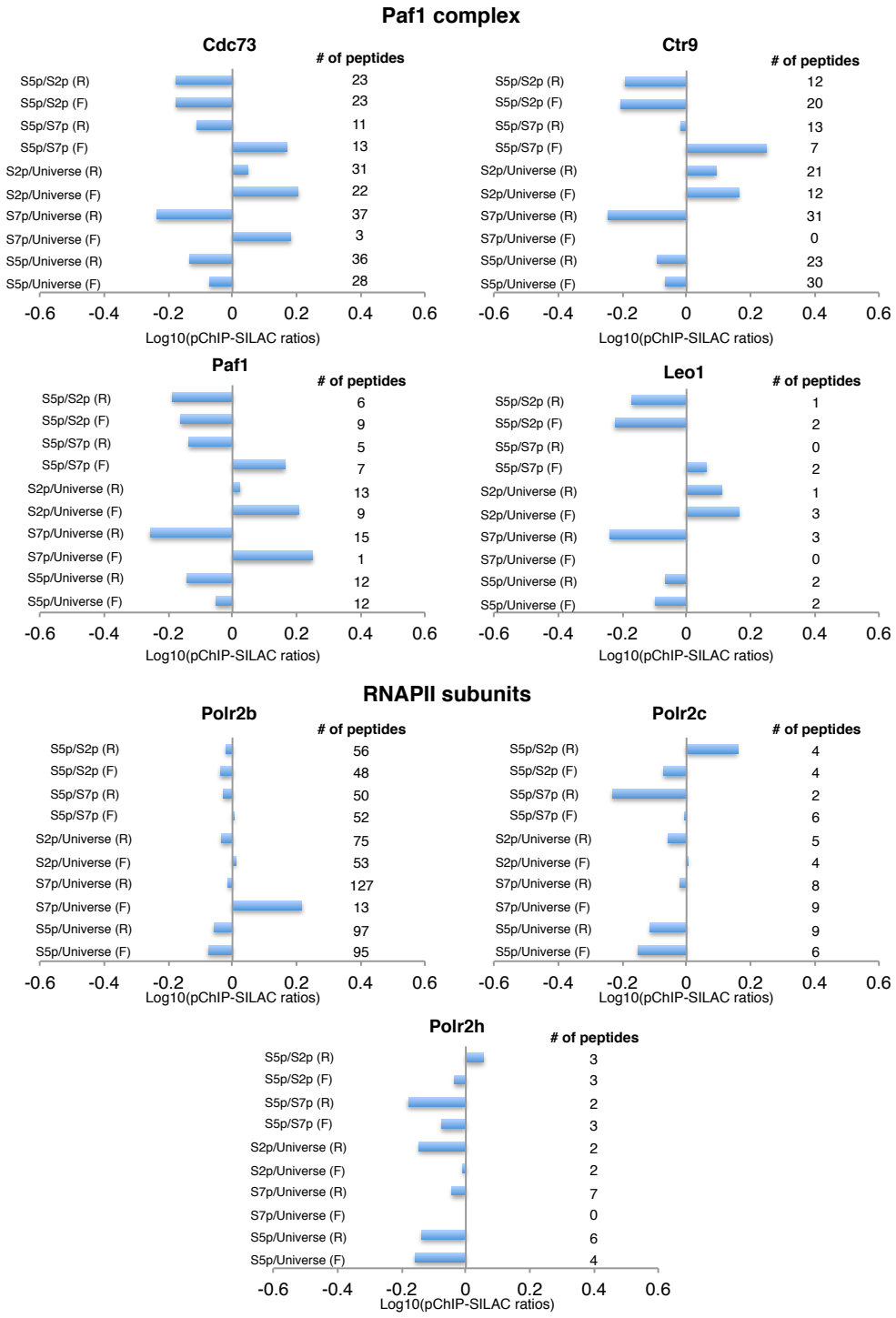


Figure 5.17. Examples of proteins with varying pChIP-SILAC ratios across 10 SILAC MS dataset. Barplot with pChIP-SILAC ratios across all 10 datasets for proteins (Paf1 complex and RNAPII subunits) that lacking any specific enrichment for RNAPII modifications. Proteins with positive log ratios are enriched heavy pChIP while proteins with negative log ratios are enriched for light pChIP. Ddx5 and Hnrnpk are involved in pre-mRNA splicing; Histone H3.2 is histone variant and Paf1 protein

is involved in transcriptional elongation and important for stem cell pluripotency. pChIP-SILAC ratios are plotted in log scale and number of peptides are represented next to ratios on barplot.

5.1.1. Comparison between different normalizations

As briefly mentioned in section 5.3.5.2, we also explored other normalization strategy to analyse our pChIP SILAC data to identify dependencies to RNAPII modifications. Rpb1 and Rpb2 were the most robustly detected subunits of RNAPII in all pChIP-MS runs and with highest peptide count, as expected from their larger size. Additionally, as each residue of the Rpb1-CTD is a site of dynamic PTM, we therefore explored if normalization with average of Rpb1 and Rpb2 pChIP-SILAC ratio was more robust than with Rpb1 ratio alone. The regression analysis was performed by Borislav Vangelov and Prof. Mauricio Barahona. All proteins and their pChIP-SILAC ratios were plotted on regression plots and the average distance of Rpb1, Rpb2, average (Rpb1, Rpb2) and all subunits was measured to the regression line using two different measures including Total Least Squares (TLS) and Random Sample Consensus (RANSAC) as highlighted in Table 5.2.

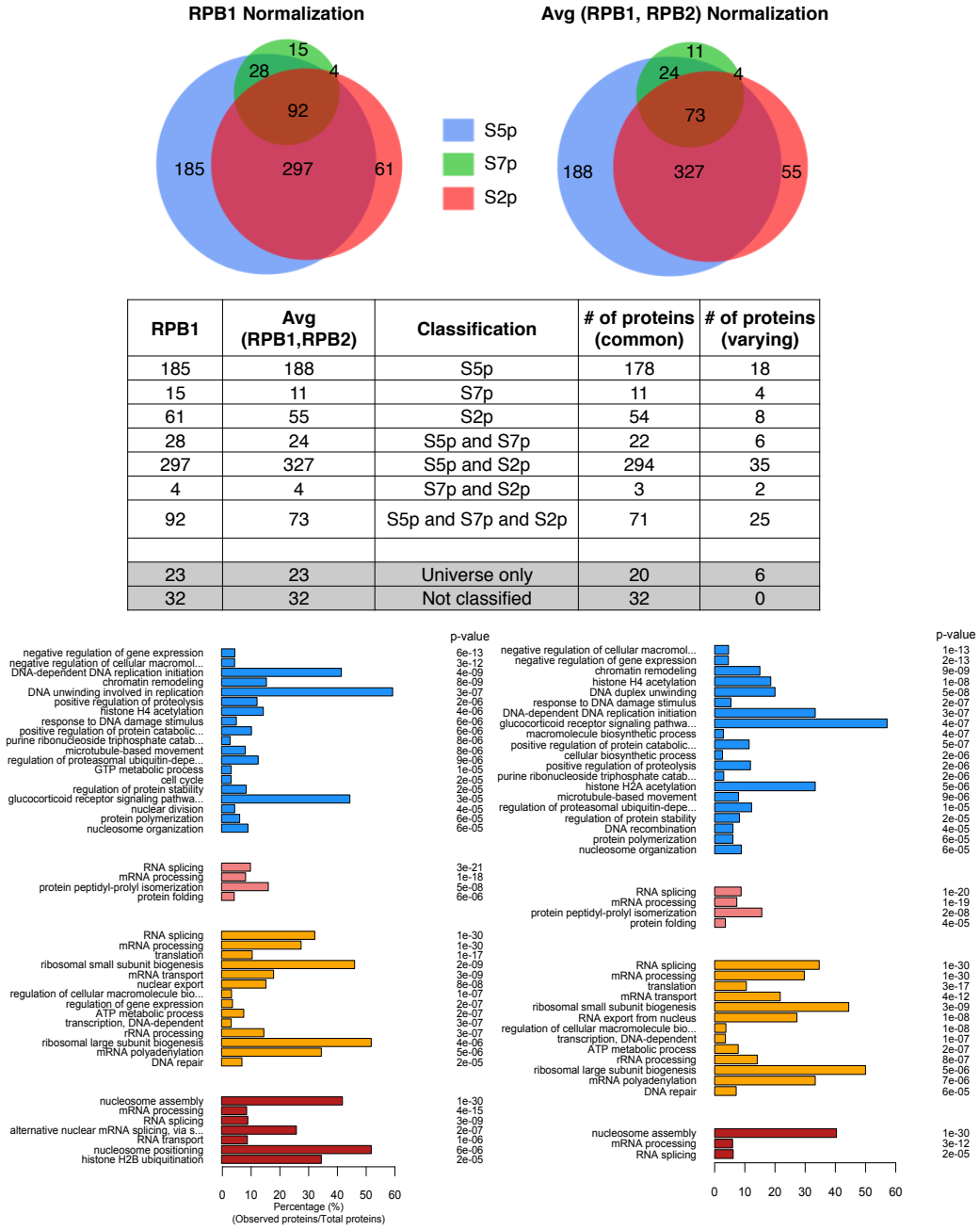
Table 5.2. Regression analysis and distances of Rpb1 and Rpb2 ratios. Distance of Rpb1, Rpb2 and Average (Rpb1, Rpb2) was calculated from the regression line. Two methods were used including Total Least Square (TLS) and Random Sample consensus (RANSAC). Analyses performed by Borislav Vangelov and Prof. Mauricio Barahona.

Method	RPB1	RPB2	RPB1 + RPB2	all subunits
TLS	0.35	0.25	0.29	0.36
RANSAC	0.26	0.19	0.18	0.26

We observed the lowest RANSAC distance compared to all subunits average (Rpb1, Rpb2) ratio. Subsequently, I also measured the effect of normalising with Average (Rpb1, Rpb2) instead of Rpb1 only, by repeating the data classification steps according to Fig. 5.4 and compared the results (Fig. 5.18). Reassuringly, similar distribution of proteins associating with specific or all

RNAPII marks was observed, with major shifts being observed for proteins moving between S5p&S7p&s2p (all modifications) to S5p&S2p (common). We observed 73 proteins associating with all RNAPII modifications (92 in Rpb1 normalization) while 327 proteins were associating with S5p&S2p (297 in Rpb1 normalization). Remaining protein groups were relatively unchanged.

S5p proteins were remarkably consistent between both normalizations (178 common out or 185 in Rpb1 and of 188 in average normalisations). Performing GO analysis, we observed almost all similar terms in all proteins groups, however with slightly different p-values and percentage of proteins in each GO category (Fig. 5.18).



RPB1	Avg (RPB1,RPB2)	Classification	# of proteins (common)	# of proteins (varying)
185	188	S5p	178	18
15	11	S7p	11	4
61	55	S2p	54	8
28	24	S5p and S7p	22	6
297	327	S5p and S2p	294	35
4	4	S7p and S2p	3	2
92	73	S5p and S7p and S2p	71	25
23	23	Universe only	20	6
32	32	Not classified	32	0

Figure 5.18. Different normalisation of pChIP ratios have minor effects in protein dependencies with S5p, S7p and S2p modifications. Normalization was done with Rpb1 ratio as before and with average (Rpb1, Rpb2) and proteins dependencies are highlighted in Venn diagram. Table representing number of proteins in both normalizations with majority of proteins common between different normalizations. Significant GO terms for both normalizations are very similar.

5.4.17. Summary of the simple binary classification analysis

To summarise the protein dependencies captured by our comprehensive and redundant experimental setup using pChIP-SILAC ratios, we schematically represented proteins and their associations to different RNAPII modifications (Fig. 5.19). Briefly, we separated active genes from genes containing only RNAPII-S5p (as described at Polycomb targets; (Brookes *et al.* 2012)) and highlighted proteins based on association with specific RNAPII modifications.

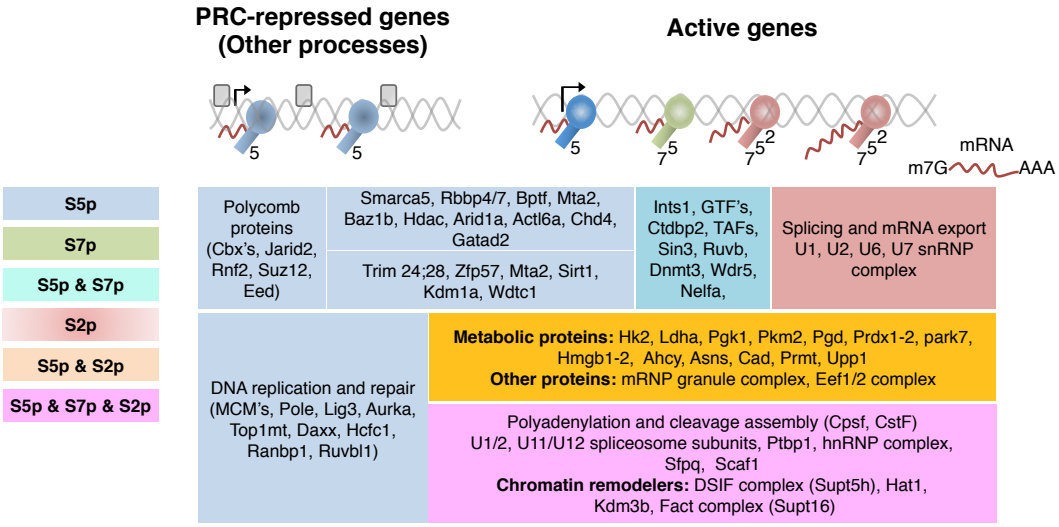


Figure 5.19. Summary of proteins and their association with different RNAPII modifications. Schematic representation of proteins and their association to RNAPII modifications during active transcription (left) and proteins with non-transcriptional biological functions.

5.5. Discussion

Genome-wide ChIP-Seq analysis from our lab has demonstrated the RNAPII-Polycomb in mES cells chromatin and have identified genes with distinct RNAPII conformation ($S5p^{+}S7p^{-}S2p^{-}$) encoding important developmental regulators, metabolic genes and several signalling pathways in a distinctly different conformation to active genes ($S5p^{+}S7p^{+}S2p^{+}$) (Brookes *et al.* 2012). Before performing pChIP, we obtained the most conservative estimate of distinct RNAPII states by re-plotting the ChIP-Seq profiles at actively transcribing genes (30% of Refseq genes) and PRC-repressed genes (10% of

Refseq genes). Given this information, I hypothesised that pChIP proteins observed on chromatin co-associating with all RNAPII modifications (S5p&S7p&S2p) would be representative of at least 30% genes (actively transcribing) of Refseq genome. In addition, we should be able to detect Polycomb proteins specifically co-existing with S5p only (without S7p and S2p). Moreover, we aimed to identify if other S5p only protein cohorts existed and their biological function.

5.5.1. Proteome-ChIP identifies large cohorts of RNAPII-bound proteins on chromatin

Our comprehensive pChIP experiment containing inherent replicates has identified >700 proteins associating on chromatin along with RNAPII modifications. We have identified the largest cohort of RNAPII associating proteins and include several known interactors and many novel interactions (Das *et al.* 2007; Jeronimo *et al.* 2007; Moller *et al.* 2012). Remarkably, our dataset also identifies several novel and uncharacterised proteins that further extend the repertoire and knowledge of RNAPII interactome. Our unbiased pChIP approach sheds light on the dynamics of interactions and further elucidates the complex RNAPII regulation on chromatin and such a systematic process has not been performed before. Our aim of elucidating the RNAPII proteome is technically quite challenging as we apply pChIP to identify and unravel chromatin proteins associated with three distinct post-translational modifications on a single protein (Rpb1) and next to each other on the Rpb1-CTD region that consists of multi-heptad repeats (instead of 2 different proteins). This task has been aided by our knowledge and use of highly specific RNAPII antibodies and their genome-wide ChIP-Seq profiles.

5.5.2. Universe approach and Pairwise approach

The universe and pairwise approaches are complementary to dissect the proteome associated on chromatin bound by RNAPII differently modified. Standalone, both universe and pairwise approach have their advantages and

caveats. While universe approach can unravel dependencies across three RNAPII modifications in a single experiment and MS run, it suffers from higher sample complexity leading to lower identification and coverage of proteins (also limited by MS depth and run-time issues). The pairwise approach on the other hand has lower complexity with better protein identification, but however it cannot identify dependencies for more than two modifications. We have demonstrated that most proteins identified in universe and pairwise experiments are complimentary to each other in identifying of proteins and their dependencies relative to RNAPII modifications.

To compare the chromatin-bound proteome between two different proteins, one could either apply just the universe approach or the pairwise approaches to reduce overall cost of experiment, but still obtain basic and simple information on candidates. However, combining the complimentary strategies adds additional layers of information towards selecting candidate interactors for biological validation. In summary, we have robustly demonstrated that pChIP using complimentary approaches robustly and succinctly identifies chromatin-bound protein cohorts.

5.5.3. Proteins co-existing with S5p only.

Our unbiased simple binary classification robustly identifies several protein cohorts that specifically associate with only S5p (no S7p or no S2p). Remarkably, very little has been reported so far about these interactions with RNAPII-S5p in the literature. Our laboratory has previously reported the RNAPII-S5p and Polycomb interplay and confirmed their co-association on chromatin by sequential-ChIP for a small number of genomic regions (Brookes *et al.* 2012). The identification of specific chromatin remodellers with S5p raises an intriguing possibility about whether additional genes (along with PRC-repressed genes) are kept in this unusual state and what factors co-associate at these regions. Further questions can also be asked about the role of RNAPII and chromatin remodellers in maintaining chromatin

architecture and genome plasticity in mES cells. Identification of DNA replication proteins with S5p-only is not only fascinating, but it also raises a range of possibilities especially in mES cells. Comparing the distribution of replication origins (Ori) in mES cells, it is striking to observe that profile of S5p completely mirror distribution of Ori observed across average gene promoters . (Cayrou *et al.* 2012). In addition, genome-wide distribution of replication timing in mES cells also provides further cues to a link with RNAPII (Hiratani *et al.* 2008). Our observation that all Mcm2-7 complex subunits were consistently enriched for S5p and depleted for S7p and S2p in both universe and pairwise experiments (Fig. 5.15), adds an interesting layer to the long-standing relationship between replication and transcription. Over several decades now, a relationship had been observed between early origins of replication in mammals and gene activity (Hassan *et al.* 1994; de Jong *et al.* 1996; Gilbert 2002; Hiratani *et al.* 2009). However this relationship seems to be distinct in ES cells, where many silent development regulator genes, repressed by Polycomb proteins, were found to be early replicating(Azuara *et al.* 2006). The discovery that Polycomb target genes are associated with RNAPII-S5p and the striking and robust detection of replication proteins in our pChIP experiments specifically associated with RNAPII-S5p sheds new light into a relationship between replication and RNAPII-S5p and not necessarily with productive transcription; in the pChIP experiments, active (S7p+S2p+) RNAPII complexes are clearly not correlated with co-association with replication proteins.

5.5.4. Proteins co-existing with combinations of RNAPII modifications.

We have demonstrated that our pChIP results robustly identify known and novel components of the transcription machinery. From our analysis, >400 proteins are associated with different combinations of RNAPII modifications. The majority of these proteins belong to S5p&S7p&S2p and S5p&S2p and are involved in various transcriptional and co-transcriptional processes. We identify many major components of the spliceosome, splicing factors and RNA

processing machinery. In addition, quite a few metabolic proteins are associated with specific combinations of modification (S5p and S2p), suggesting additional roles for these proteins and importantly the balance between mES cell metabolism, energy demands and transcription control (Brinster and Harstad 1977; Chen *et al.* 2012). We also identify proteins with GO terms related with protein synthesis, including ribosome biogenesis and translation. This result is consistent with the concept of nuclear transcription-coupled-translation, a topic which has been debated in past, but more recently further reinforced (Iborra *et al.* 2001; Dahlberg *et al.* 2003; Iborra *et al.* 2004; David *et al.* 2012).

The identification of S2p only proteins from the pChIP dataset also raises interesting insights in our ability to capture simultaneous modifications on the Rpb1-CTD. S5p and S2p are found to co-exist through the coding regions of active genes (genome-wide ChIP-Seq), and as measured for some genes at the same chromatin regions simultaneously (by sequential-ChIP;(Brookes *et al.* 2012). These results raise the possibility that some genomic regions, or particular transcriptional states are either associated with RNAPII-S5p^{-}S2p^{+}, or may be associated with RNAPII-S5p^{+}S2p^{+} on RNAPII complexes where the S5p antibody (4H8; Table 2.1) does not bind. Interestingly, ELISA results from the laboratories of Dirk Eick and Hiroshi Kimura (referred as 'not shown' in (Brookes *et al.* 2012), show that S5p in the Rpb1-CTD heptad (**C** Y_1-S_2-P_3-T_4-S_5-P_6-S_7 **N** terminal) can successfully detect **C**-S5p-S2p-**N** peptides (with 3 unmodified amino acid residues in between), but not **C**-S2p-S5p-**N** residues with 2 residues in between; in the case of the H5 antibody, used to immunoprecipitate S2p, it is not affected by the double modification **C**-S2pS5p-**N**. It is interesting to speculate whether the S2p-only proteins may correspond to heavily phosphorylated CTD repeats, where the 4H8 antibody fails to access. To this date, of all S5p antibodies tested by ELISA, 4H8 is the one least affected by other modifications, so it is still the option of choice, but

future experiments with novel S5p antibodies designed to capture more complex phosphorylation patterns would shed further light on this topic.

5.5.5. Rpb subunits and normalization.

Our pChIP experiments consistently and robustly detect both Rpb1 and Rpb2 across all 10 MS dataset. We have demonstrated that normalization to Rpb1 or Rpb2 or average (Rpb1, Rpb2) does not affect the overall dependencies of protein to RNAPII modifications. However, there is a minor re-arrangement of proteins between groups. We suspect that these proteins are not consistently enriched in a particular modification or are not always detected across most experiments and therefore are susceptible to normalization. We also observed proteins that do not have specific enrichment for RNAPII modifications (Fog. 5.17; Paf1 complex and others) and most of these proteins have inconsistent ratios between replicate pairs (forward and reverse). The pattern of pChIP ratios for some of these proteins are extremely similar to Rpb subunits (Fig. 5.8; lie in the vicinity of Rpb1) and therefore normalization affects their pChIP-SILAC ratios. We can only perform limited visual confirmation for these varying ratios, and in additional only detect patterns which are compatible with our a priori knowledge of RNAPII regulation and chromatin interactome. Therefore, an additional and unbiased systems biology approach is required to unravel novel patterns (including stoichiometry of interactions) inherent to our pChIP MS dataset that cannot be resolved from visual observation or simple classification.

6. Unravelling the network landscape of RNAPII interactome

6.1. Research motivation

The chromatin-bound RNAPII proteome and its landscape is quite diverse in mES cells. In Chapter 5, I describe a comprehensive pChIP experiment and a simple classification that unravels proteins cohorts and dependencies on specific RNAPII modifications. Moreover, our pChIP-SILAC ratios encompass information on stoichiometry of interactions, relative affinity and strength of interaction on chromatin relative to Rpb1. My aim in this chapter was to apply an unbiased machine learning approach to sensitively explore, detect patterns of proteins enrichment for RNAPII modifications that cannot be unravelled by visual or simple classification. We aimed to perform integrative systems biology analysis on pChIP-SILAC ratios to efficiently group the proteins based on pChIP-SILAC ratios relative to RNAPII and generate a protein network to visualise the patterns.

Borislav Vangelov and Prof. Mauricio Barahona performed all mathematical analyses in an active collaboration between our laboratories. I performed the GO analyses (along with other auxiliary analyses) and interpretation was achieved in collaboration between our two laboratories.

6.2. Results

Our pChIP comprehensive experiment which includes 12 datasets (Fig. 5.2) was carefully devised to contain replicate information and to capture both the transient and more robust protein associations occurring on chromatin along with distinct RNAPII modifications, S5p, S7p and S2p. Using a simple classification system based only on the direction of SILAC ratios (positive or negative; Fig. 5.12), we have highlighted and began to dissect the interesting and specific cohorts of proteins that bind to chromatin with RNAPII modifications; this led us to unravel interesting and, in some cases, unexpected cohorts of proteins that co-associate with RNAPII-S5p on chromatin. However, the information encoded in the SILAC ratios is much deeper including the magnitude of the ratios, which bears information on stoichiometry of protein co-association with chromatin, which will depend on whether protein associations are inter-dependent (for example for subunits of the same protein complex) or not (for complexes that may independently associate with chromatin-bound by the same RNAPII variant, for example, on different genomic regions. Therefore we decided to use an unbiased approach, based on machine learning to unravel patterns of protein co-association with RNAPII-bound chromatin, and systematically dissect protein associations that could not be unravelled by visual inspection of the complex pChIP datasets. We used a novel clustering approach, Markov stability for community detection (Delvenne *et al.* 2010). Vangelov and Barahona have further developed this approach through the construction of a network from the similarity/distance matrix by using a perturbative minimum spanning tree (PMST) method (B. Vangelov and M. Barahona, in preparation). These novel clustering and network approaches have recently been applied on large and complex networks characterized by the presence of communities of communities (multiscale communities; (Schaub *et al.* 2012) or locally sparser networks where not all components are connected to each other but instead locally related. This could be the case in our global analyses of proteome

associated genome-wide with RNAPII, as we will have proteins that coincide on chromatin on the same genes but not others, or during a given cell cycle stage but not another. If two proteins (or protein complexes) are functionally related, the magnitude of their SILAC enrichment (a measure of stoichiometry) is more likely to be well correlated across all datasets, than if the association of the two proteins with chromatin characterized by a given RNAPII modification occurs on separate genomic regions or cell cycle stages. In this scenario, subunits of the same protein complexes should show the strongest correlation between SILAC enrichments across the ten-pChIP datasets.

6.2.1. Summary of network and clustering analysis

For analysis of pChIP-SILAC ratios, we used Markov stability for community detection (Delvenne *et al.* 2010), followed by the construction of a network from the similarity/distance matrix by using a perturbative minimum spanning tree (PMST) method (B. Vangelov and M. Barahona, in preparation) following the steps outlined in Fig. 6.1; data imputation, network, clustering and network property analyses were done by Borislav Vangelov.

The set of ten pChIP-SILAC MS datasets is composed of a list of all proteins detected in at least one of the MS datasets. As highlighted in Chapter 5 (Figs. 5.13 – 5.17), some proteins are detected in all MS datasets (e.g. chromatin remodellers and Paf1 complex; Figs. 5.15 and 5.17), whereas other proteins are identified in only some datasets (Figs. 5.14 and 5.16). Missing values are not allowed in the networking analysis, and require imputation.

Imputation is a process of substituting missing data with a probable value estimated from the other available information, in our case from the relationship between ratios available across datasets. To explore the effect of imputation, we started by considering different sub-lists of proteins from the total list, which contained proteins with fewer or larger number of missing values (requiring lower or higher extent of imputation, respectively):

A. List of 446 proteins identified in a minimum of five pChIP MS datasets (out of 10), with at least one value in the Forward or the Reverse dataset for a given pChIP.

B. List of 702 proteins (including the above 446 proteins) identified in minimum of three pChIP MS datasets (out of 10), with at least one value in the Forward or the Reverse dataset for a given pChIP.

C. List of 858 proteins, identified in a subsequent re-analysis of the pChIP MS spectra.

In this chapter, I will first describe the analyses of the dataset with 702 proteins. The comparison between the first two lists (446 and 702 proteins) gave similar results and is described at the end of this chapter. The analyses of the larger dataset (858 proteins) also produced similar results, but required extensive imputation, and for this reason was not followed further and is not shown or discussed further in this thesis.

Before performing the data analysis, we first looked at individual experiments and how well proteins were detected across the ten-pChIP experiments as previously, mock/Universe (F) and (R) datasets were used to identify contaminants. Since data imputation is required before clustering to yield a dataset without missing values, we inspected the extent of protein detection across each dataset. We observed that most proteins were robustly detected in all experiments except ‘S7p/Universe (F)’ experiment where only 21% of protein ratios were observed. To minimise the imputation of ratios, we chose to exclude this dataset from the following steps of clustering and network analyses; in this process we loose two proteins only detected in this dataset, resulting in a list of proteins with 700 proteins, instead of 702. The analysis pipeline is summarized in Fig. 6.1.

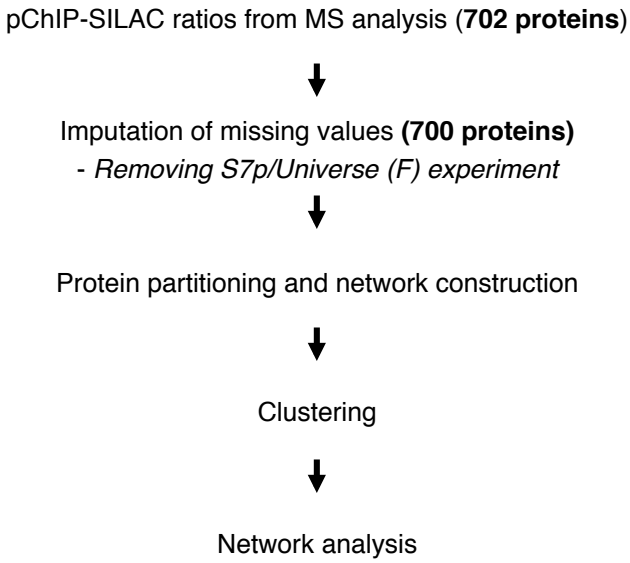


Figure 6.1. Data analysis pipeline involved in clustering and network analysis. pChIP-SILAC dataset generated by MS analysis comprehensive experimental setup (as described in Fig. 5.2) was used for clustering and network analysis. The pChIP experiments and values for 700 proteins (after removing S7p/Universe experiment) were imputed to estimate missing values.

6.2.2. Data Imputation

The SILAC ratios are traditionally represented as the ratio between heavy and light peptides. To facilitate visualisation and comparison of the pChIP ratios between Forward and Reverse datasets in the same kind of pChIP experiment, we first inverted the reverse experiment ratios. Imputation for the pChIP-SILAC datasets was performed using k-nearest neighbour imputation (Cover 1982). We used a weighted Euclidean distance to find the nearest neighbour (Fig. 6.2).

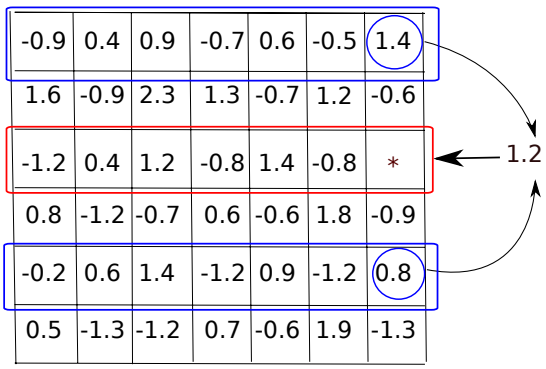


Figure 6.2. Example of missing value imputation by k-nearest neighbour. Weighted Euclidean distance was calculated for all proteins in the nine-pChIP experiments considered (excluding S7p.U_F) to find the k-nearest neighbour and to impute the missing value.

To test whether the imputation process could have added many outliers to the pChIP-SILAC ratio dataset, we measured the distribution of pChIP-SILAC ratios before and after imputation for all pChIP-SILAC forward and reverse experiments. The distribution of pChIP-SILAC ratios was maintained after imputation and in addition the median values were relatively unchanged. The Reverse and Forward pChIP datasets after imputation were treated as separate pChIP experiments, as opposed to being averaged to keep our analysis unbiased and to observe any variation due to imputation.

6.2.3. Protein grouping, network construction and partitioning.

Protein ratios across the 10-pChIP datasets are represented in a nine-dimensional space. Each point in the multi-dimensional space corresponds to a single protein and connections (*i.e.* edges) between the proteins are calculated based on pChIP ratios by Clustering and Network algorithm Delvenne et al. PNAS; (Schaub *et al.* 2012); B. Vangelov and M. Barahona, in preparation. Briefly, our algorithm identifies at least one connection between each pair of proteins in the multi-dimensional space and iteratively identifies connections for all. This is in contrast with other clustering methods where

distances between all proteins against each other are computed in single step. During our iterative algorithm, protein ratios are perturbed in multiple iterations and connections are overlaid in multi-dimensional space. The union of these multiple iterations forms the minimum skeleton base network of connections between all proteins.

The rationale for partitioning of network follows the pattern of pChIP-SILAC ratios for proteins across nine MS datasets. Briefly, proteins that have similar pattern of distribution of pChIP-SILAC ratios across the nine different MS datasets proteins are well connected with shorter distances. In addition, they are preferentially placed close to each other in multi-dimensional space. Whereas proteins with dissimilar pattern have larger distances and are more separated.

'Stability' is additional mathematical parameter applied to partitioning of network and robustly reinforces the identification of partitions. Briefly, 'stability' measures the network architecture identifying the optimal transitions between clusters and computing number of stable clusters from the network itself.

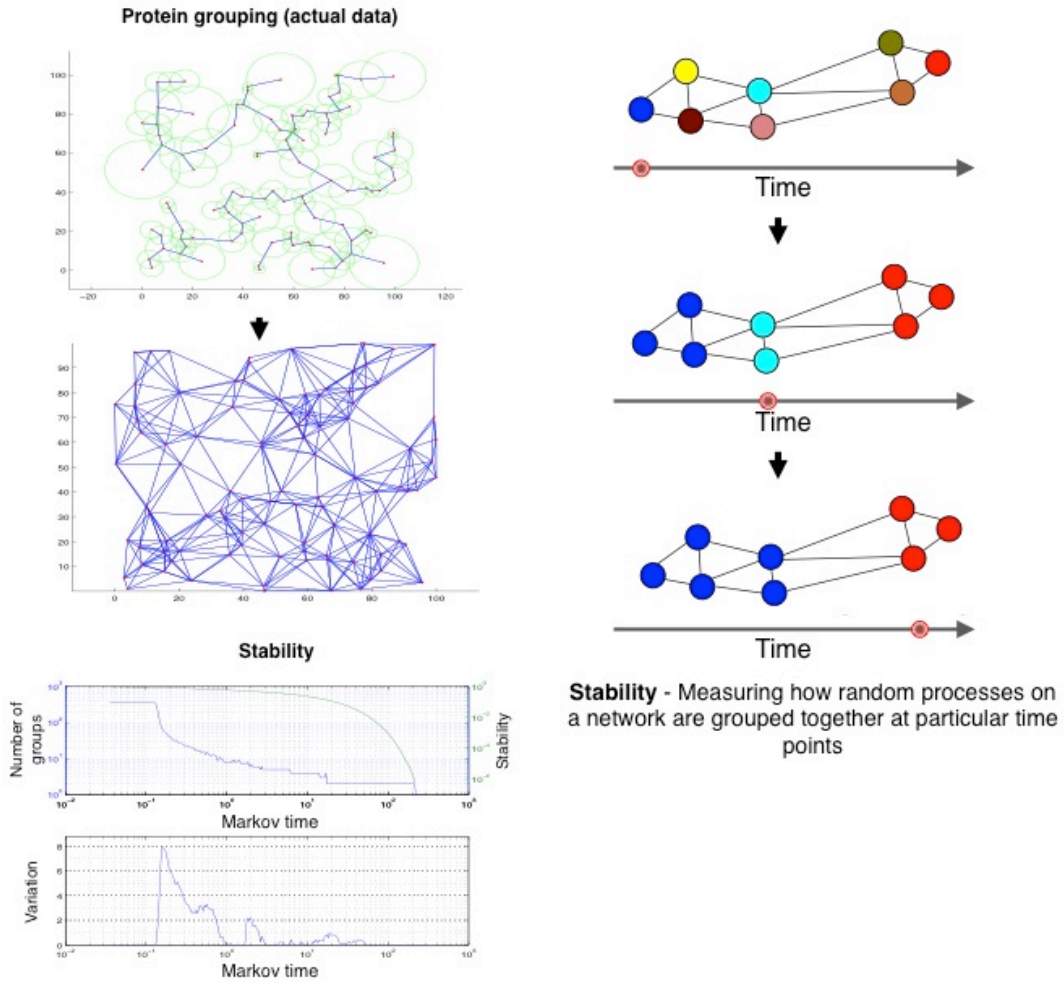


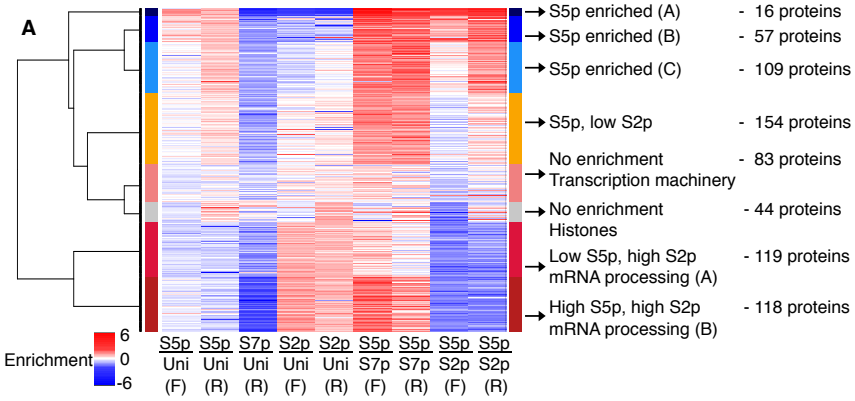
Figure 6.3. Simplistic representation of protein grouping, partitioning and network construction. Protein (with pChIP-SILAC ratios) are plotted in nine-dimensional space (9 experiments) and iteratively perturbed to construct a base skeleton for the network. The stability parameter is calculated by an algorithm and identifies robust number of clusters by examining the transitions between clusters in base skeleton (Bottom left graph; number of breaks).

6.2.4. Clustering of protein groups

Clustering and network analysis of the pChIP datasets identified eight clusters as most relevant and robust, which can be visualised in the form of a hierarchical tree (Fig. 6.4A). Each of the nine pChIP experiments form the columns of the clustering tree with intensities across each row representing the pChIP-SILAC ratio for each protein across the nine pChIP experiments (Fig. 6.4A; following inversion for the Reverse datasets, for simplicity). The length of the dendrogram represents the robustness in separation of clusters

as distance metric, *i.e.* the longer the length of dendrogram, the more robust and dissimilar are the clusters.

Visual inspection of the pChIP SILAC ratios across the eight clusters derived after clustering shows that they contain groups of proteins characterized by specific levels of SILAC enrichment relative to specific RNAPII modifications. Proteins in the top three clusters (Clusters 1- 3; dark blue, blue and light blue) are highly enriched in the S5p pChIP's only (with low or no enrichment for S7p and S2p). Cluster 4 contains proteins enriched for S5p and at low levels for S2p (Orange). Proteins in Clusters 5 and 6 had basal level of enrichment for all three RNAPII marks (pink and grey). Proteins in clusters 7 and 8 have high enrichment for S2p with varying enrichments for S5p. Numbers of proteins in each cluster are also represented in Fig. 6.4A. A simplistic table with cluster names and RNAPII enrichments is described in Fig. 6.4B. Some examples of proteins in each cluster are highlighted in Fig. 6.4C (and also described in detail below) and significant GO terms for enriched proteins in different clusters is represented with cluster colours in bar-plot (Fig. 6.4D).



B

#	Cluster name	S5p	S7p	S2p
Cluster 1	S5p enriched (A)	High	-	-
Cluster 2	S5p enriched (B)	Moderate	-	-
Cluster 3	S5p enriched (C)	Moderate		
Cluster 4	S5p and low S2p	Moderate	-	Low
Cluster 5	Transcription machinery	Low	Low	Low
Cluster 6	Histones	Low	Low	Low
Cluster 7	mRNA processing (A)	Low	-	High
Cluster 8	mRNA processing (B)	Moderate	-	High

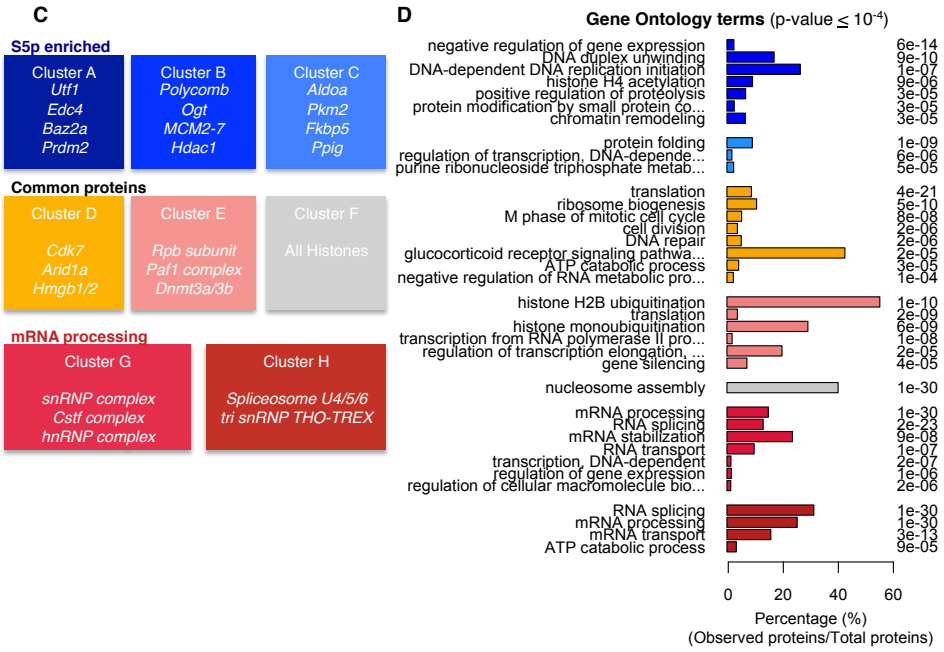


Figure 6.4. Clustering of pChIP-SILAC ratios results in eight stable, robust clusters delineating a gradient separation of proteins. (A) Clustering of pChIP-SILAC ratios resulted in robust separation of 8 stable and robust clusters with different pChIP enrichments for different RNAPII modifications. Each experiment is represented as a column in the clustering image. Cluster names were defined based on enrichment and the functionality of proteins in each clusters as identified by visual inspection of the list and by GO analyses. Intensity values (-6 to +6) represented Log scale and represent values of enrichment. (B) Simplified representation of RNAPII enrichment levels across different clusters. (C) Examples of proteins present in the

different clusters. (D) Significant GO terms for different clusters; cluster 1 did not yield any GO term enrichment as expected for its low number of proteins. Bar colours in bar-plot are representative of cluster colours.

The top three clusters enriched for S5p only were termed ‘S5p enriched (A)’, ‘S5p enriched (B)’ and ‘S5p enriched (C)’. Cluster 1 consists of 16 proteins including Utf1 (undifferentiated stem cell transcription factor), Edc4 (enhancer of decapping), Baz2a (essential component of nucleolar remodeling complex) and Prdm2 (methyltransferase). Due to small number of these proteins, significant GO terms are not observed, however the discrete clustering and enrichment levels suggest the strongest association of this small group of proteins with RNAPII-S5p only and highlights the ability of the clustering approach to identify a small group of proteins which were not yet know to be related with RNAPII modification. The presence of Utf1, an obscure pluripotency factor until recently(Jia *et al.* 2012), was particularly noteworthy, along side with ill-characterized chromatin remodelers and RNA processing component, Edc4.

Cluster 2, ‘S5p enriched (B)’, consists of 109 proteins and includes Polycomb proteins (Suz12, Ring1b, Jarid2, Rbbp7, Rbbp4), DNA replication (all MCM2-7 subunits), chromatin remodellers (Smarcc1, Mta2, Smarca5, Rbbp7), negative regulators of transcription (Hdac2, Sin3a, Kdm1a, Cobra1) and proteins involved in histone acetylation (Ruvbl1-2, Dmap1).

Cluster 3, ‘S5p enriched (C)’, consists of 57 proteins involved in protein folding (Cct2, Cct3, Cct5), transcriptional regulation (Klf5, Sirt1, Mta1, Mta3) and metabolic processes (Aldoa, Pkm2). Pkm2 works as a chromatin modifier (Yang *et al.* 2012), although previous association with RNAPII has not been highlighted. Notably, we identify three peptidyl-prolyl cis-trans isomerases (Fkbp3, Ppig, Ppid), the latter two not previously related with nuclear processes.

Cluster 4 is predominantly enriched for ribosomal proteins and metabolic proteins. These 154 proteins have low levels of enrichment relative Rpb1 in the S5p and S2p pChIP experiments (not for S7p) and include proteins Cdk7 (cyclin-dependent kinase and transcription), Arid1a (SWI-SNF protein) and Hmgb1/2 (high mobility group proteins that directly bind and bend DNA). Enriched GO terms include 'translation', 'DNA repair' and 'ATP catabolic processes'.

The central clusters, 5 and 6, consist of RNAPII subunits, TAFs (transcription associated factors) and histones. These proteins have similar enrichment relative to Rpb1 and showed not consistent preference for RNAPII modification. Cluster 5 consists of 83 proteins including 8 RNAPII subunits. Remarkably we also identify all components of the Paf1 complex, which are involved in transcription elongation, but also have important roles in regulation of development and maintenance of stem cell pluripotency (Ding *et al.* 2009). Cluster 6 consists of 44 chromatin proteins, predominantly histones and few transcription-associated-factors (TAFs). Only significant GO term is "nucleosome remodelling".

Clusters 7 and 8 have proteins with high enrichment for S2p, but varying levels of S5p and enrichment for transcription elongation proteins including mRNA processing, spliceosome subunits, splicing factors and RNA export factors.

6.2.5. Network landscape and properties

To understand the relationships between clusters and proteins in each cluster, we next constructed a network model of the partition with eight clusters. In the network model, every node represents a cluster and two nodes are connected if any two nodes from the corresponding clusters are connected. The simplistic network model of the partition with eight clusters is shown in Fig. 6.5A. The complete pChIP network with cluster names, colors and

connections is shown in Fig. 6.5B. The network was plotted with package Gephi (version 0.8.1) and the layout was done using ‘force atlas layout’.

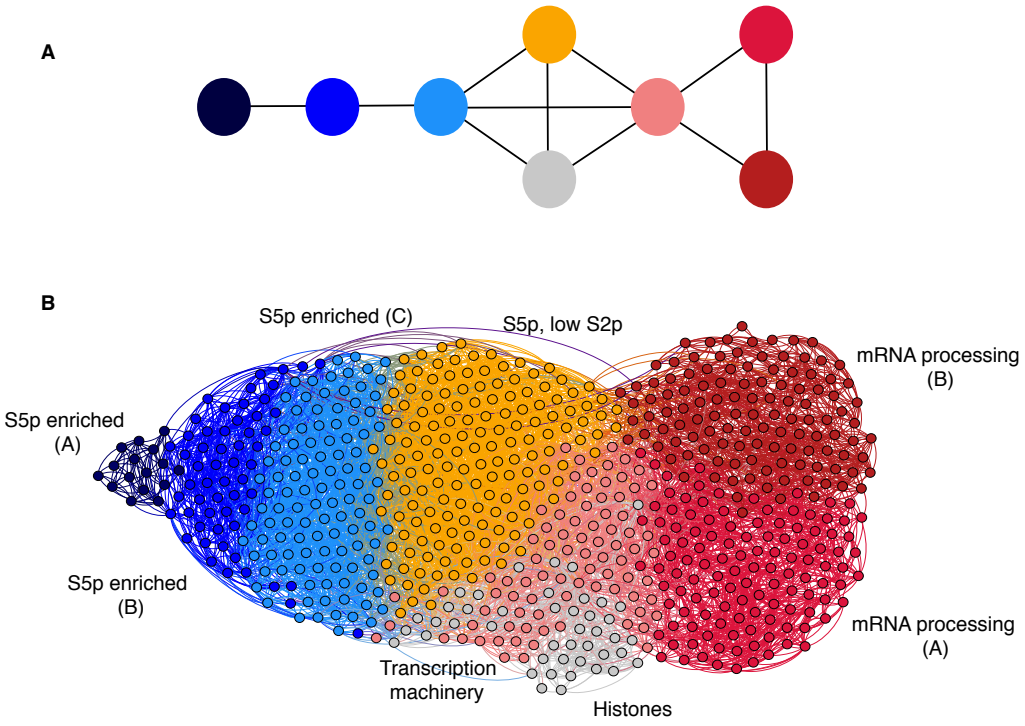


Figure 6.5. Network landscape of clustered proteins and connections. (A) Estimation of network model of a partition with each node representing a cluster and edges representing connections between clusters. (B) Visual representation of protein network based on pChIP-SILAC ratios using Gephi (version 0.8.1). Nodes represent individual proteins and edges represent distance vector between the nodes. Weight of the edges is indicative of the strength of association (similar vector distances). Cluster colours are overlaid on the proteins. Clear separation of proteins in different clusters is a first indication of robust partitioning.

Looking at the pChIP protein network characteristics, the clear separation of proteins and cluster colours is the first indication of the robustness of clustering and efficient partitioning. The high density of connections within the network indicates similar behaviour of proteins relative to RNAPII modifications (either stoichiometry or similar protein complexes) and their association. The intra-cluster connections indicate transition between clusters and highlight the importance of bridging proteins that link the clusters.

To visualise how the network structure related with the original pChIP-SILAC ratios, we averaged forward and Reverse pChIP ratios and represented the average ratio for each protein overlaid on the network separately for the five-pChIP types of experiment (Fig. 6.6). These average enrichment profiles, represented over the network structure, highlight the successful gradient separation of proteins across different clusters and proteins within specific clusters. For example, in the pairwise S5p.S2p experiment, the average plot (Fig. 6.6, far right) shows proteins enriched for S5p in the top, proteins in the middle have low levels of enrichment relative to all RNAPII modifications (have enrichment similar to Rpb1), and finally proteins towards the bottom are enriched relatively to the S2p modification. These average intensity profiles offer an alternative way to explore the protein behaviours relative to the network structure and to evaluate the quality of the clustering and network produced.

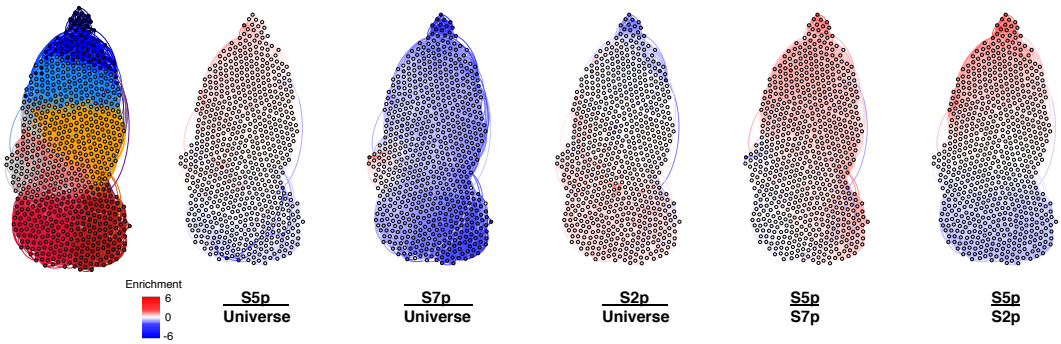


Figure 6.6. Visualising individual protein intensities across different experiments on the network structure. Average pChIP enrichments were represented for each protein on the network after averaging forward and reverse pChIP values, except for the S7p/Universe experiment for which on the Forward values were represented.

6.2.6. Important proteins in each cluster

Further exploring the properties of the pChIP network, we next zoomed in to different clusters and observed the protein connections. As the pChIP experiment was performed to identify the proteome of RNAPII-bound chromatin, we first asked how many RNAPII subunits were captured by pChIP clustering and network analyses and their positioning within the network.

Strikingly, we identify seven RNAPII subunits within the same cluster and importantly connected with each other with similar weights (Fig. 6.7, bottom right). Interestingly, other proteins such as Smc2, Ell and Dnmt3a/b are found directly connected to Rpb1 subunits, which results from similar behaviours of pChIP ratios across experiments.

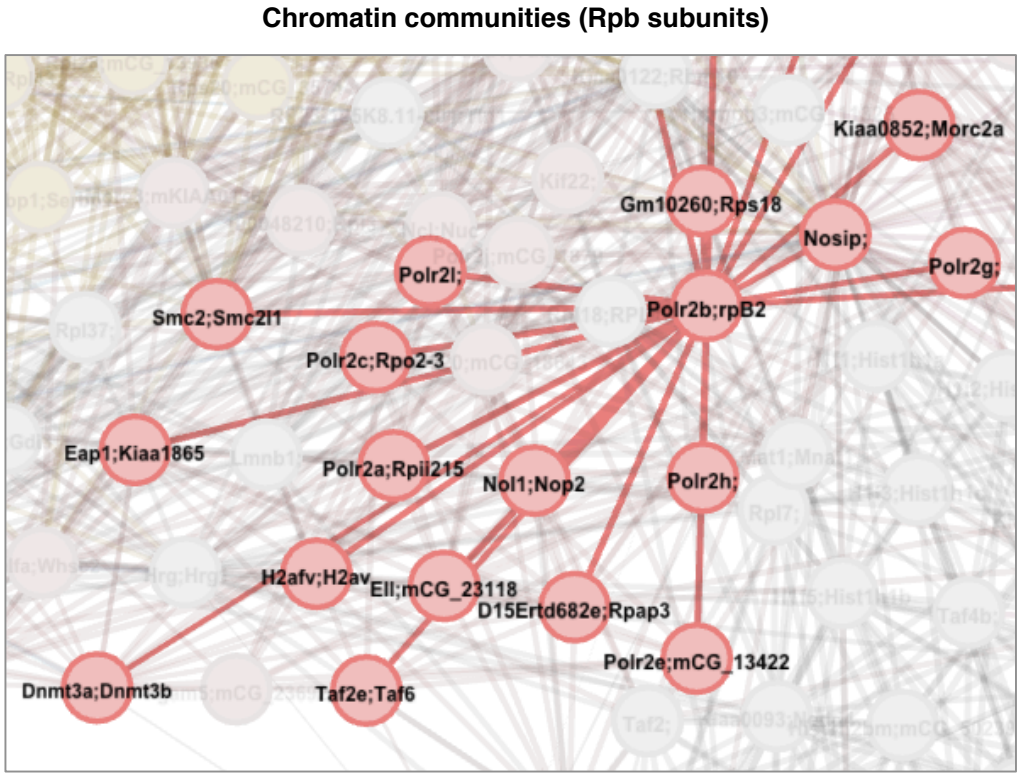


Figure 6.7. Snapshot of Rpb subunits and chromatin communities captured by pChIP. Eight RNAPII subunits (all connected within Cluster 5) are identified from pChIP network. Gephi (version 0.8.1) was used to visualise the network.

We next looked at Polycomb proteins (cluster 2) and more specifically at one of the PRC1 (Polycomb repressive complex 1) subunits, Ring1b (Fig. 6.8). Ring1b is a nuclear protein with E3 ubiquitin ligase activity and is responsible for mono-ubiquitination of lysine residue of histone H2A (H2Aub). Polycomb-repressed genes are occupied by high levels of RNAPII-S5p (without S7p or S2p) along with H2Aub (Stock et al. 2007, Brookes et al. 2012; see also Fig. 5.1 and section 5.3.1). In our pChIP network, we observe three Polycomb

proteins (Suz12, Jarid2 and Ring1b) and components of chromobox proteins (Cbx3 and Cbx5) in the same cluster. Interestingly, Ring1b protein was found directly and strongly connected to O-GlcNAc transferase subunit p110 protein (Hoffmeyer *et al.* 2012).

Ogt is the only nuclear protein known to catalyse transfer of β-N-acetylglucosamine (*O*-GlcNAc) to serine or threonine residues of target proteins. Ogt had previously been shown to co-associate with Polycomb proteins in mES cells (Myers *et al.* 2011), and has been recently shown to modify RNAPII-CTD (Ranuncolo *et al.* 2012). Interestingly, Ogt is robustly identified in our dataset co-existing only with S5p and it bridges most connections between clusters 2 and 3. Our results suggest that the functional relationship between Ogt and Polycomb, previously identified by in vitro approaches, co-exists with RNAPII-S5p (not S7p or S2p) on chromatin.

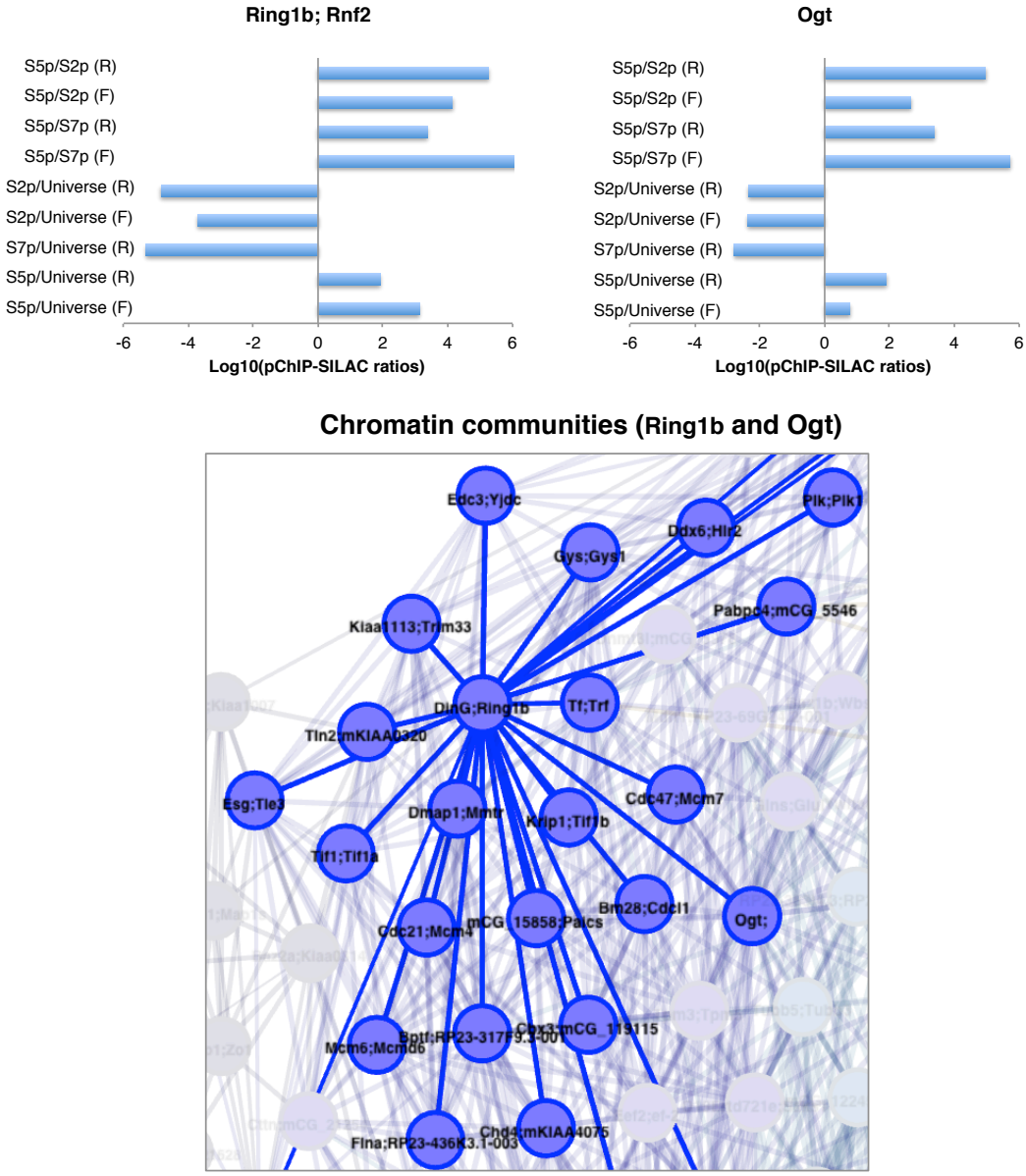


Figure 6.8. Snapshot of Ring1b and chromatin communities captured by pChIP. Ring1b (Cluster 2) are identified from pChIP network and was directly linked to Ogt protein. Gephi (version 0.8.1) was used to visualise the network. Barplots of Ring1b and Ogt proteins with pChIP-SILAC ratios after imputation. H/L SILAC ratios were inverted in the reverse (to be similar to forward) experiments and represented as indicated. pChIP-SILAC ratios are plotted in log scale and numbers of peptides are represented next to ratios on barplot.

Understanding replication control in stem cells has been of intense focus in the last couple of years with many studies, including genome-wide datasets, shedding light on need for dynamic and fast replication in stem cells and

coordinated genome integrity (Azuara *et al.* 2006; Hiratani *et al.* 2008; Sequeira-Mendes *et al.* 2009; Cayrou *et al.* 2012). In our pChIP-network, we identify all subunits of MCM2-7 complex, remarkably within the same cluster with association on chromatin with S5p only (no S7p or S2p) and with robust connections between themselves. Our finding that RNAPII-S5p specifically co-exists on chromatin along with DNA replication components raises an interesting viewpoint on dynamic between replication and RNAPII coordination specifically from the context of mES cell biology.

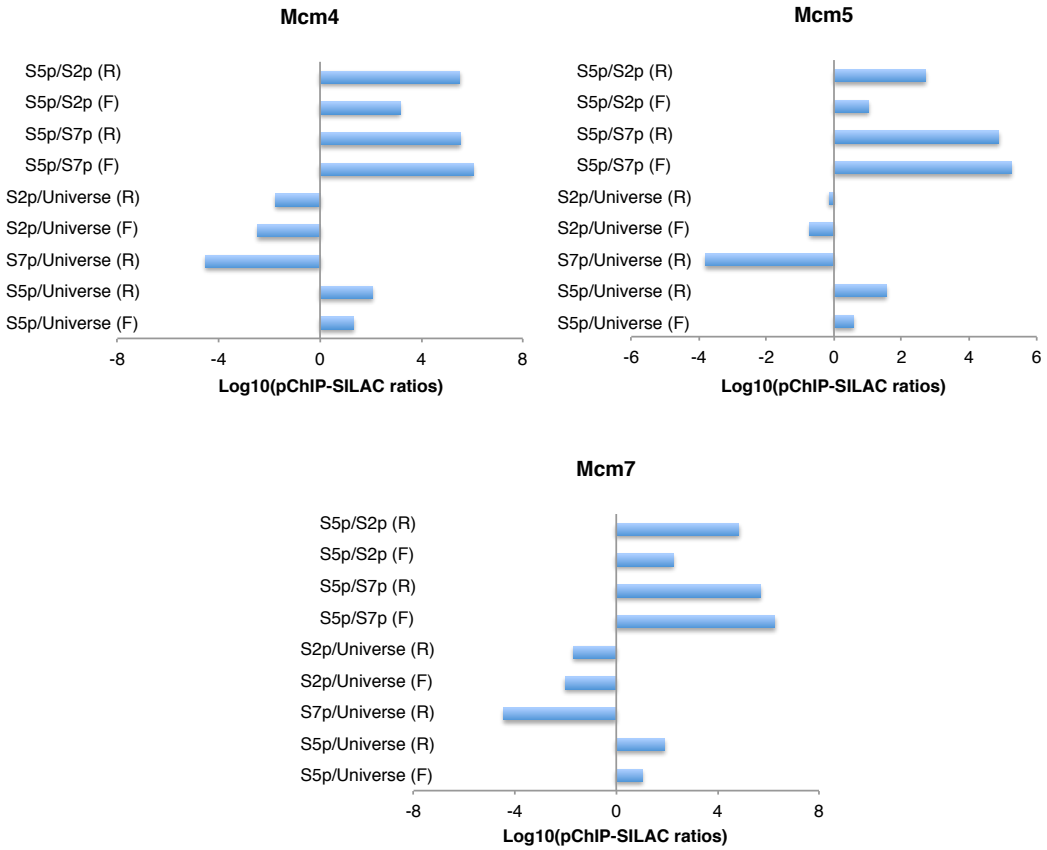


Figure 6.9. Consistent detection of DNA replication proteins with S5p only (after imputation). Barplot for three MCM proteins (Mcm4, 5 and 7) and their pChIP-SILAC ratios highlight specific association on chromatin with RNAPII-S5p. H/L SILAC ratios were inverted in the reverse (to be similar to forward) experiments and represented as indicated. pChIP-SILAC ratios are plotted in log scale and numbers of peptides are represented next to ratios on barplot.

We next looked at cluster 1 that consisted of only 16 proteins and were specifically and mostly enriched for S5p only (without S7p or S2p; Fig. 6.8). It is remarkable that none of these proteins has been shown in the literature to associate with RNAPII and in our experiment these proteins were more robust S5p associations on chromatin than Polycomb protein or other known processes (such as splicing factors with S5p&S2p). This result highlights the power of an unbiased systems approach to uncover novel regulators of RNAPII, chromatin and RNA processing through pChIP experiments using RNAPII immunoprecipitation. Edc4 is a component of the mRNA decapping and degradation-based post-transcriptional gene silencing (PTGS) complex. It is thought to associate and enhance the activity of Dcp2 in decapping the m7G cap from the mRNA (Tritschler *et al.* 2009). Baz2a is essential component of the NoRC (nucleolar remodeling complex) complex and mediates silencing rDNA by recruiting chromatin remodelers and DNA methyltransferases, leading to heterochromatin formation and transcriptional silencing (Uniprot 2012). An association with RNAPII had not been previously uncovered. Prdm2, a 'S-adenosyl-L-methionine-dependent histone methyltransferase', specifically methylates H3K9 in humans and may interplay with RNAPII in setting up heterochromatin formation; this association was not previously reported (Wu *et al.* 2010).

Utf1 is an essential stem cell transcription factor, which may associate with the TFIID complex via TBP-mediated interactions in mES cells (Fukushima *et al.* 1998). More recently, Utf1 has been related with silencing of PRC-repressed genes through RNA pruning and association with decapping proteins (Edc3/Edc4; (Jia *et al.* 2012). This study has failed to identify an association with RNAPII-S5p on chromatin probably due to the use of non-chromatin extracts in non-native *in vitro* conditions. Future analyses of Utf1 co-association with RNAPII-S5p occupied chromatin by sequential-ChIP at panel of few genes would provide further evidence for this association, but requires knowledge of the genomic regions where this association may exist.

Bioinformatic analyses of genome-wide maps of Utf1 and RNAPII-S5p would identify candidate regions occupied by both markers. Currently, available commercial antibodies for Utf1 are not ChIP grade and work only with western blotting.

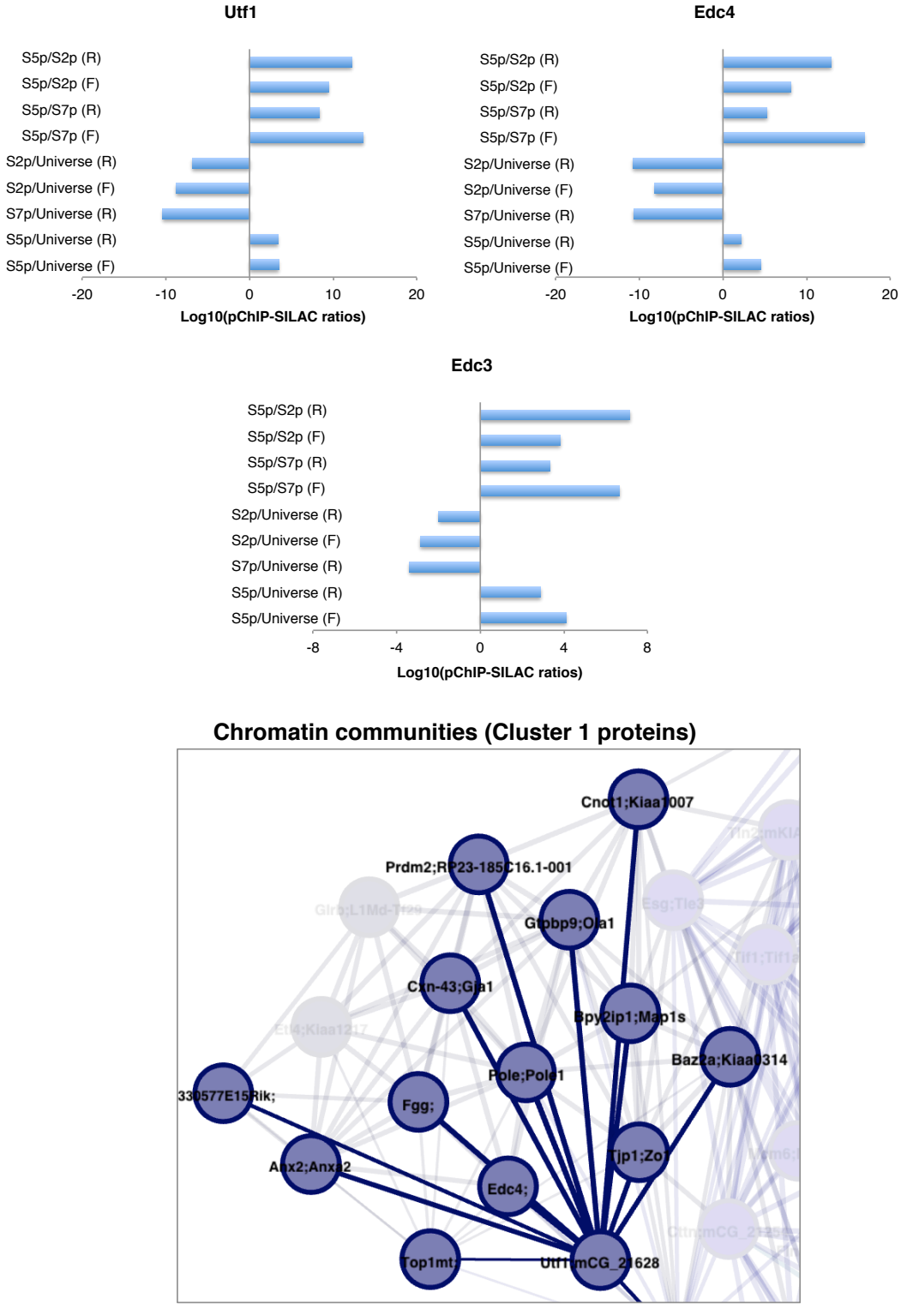


Figure 6.9. Snapshot of proteins in cluster 1 and representative chromatin communities captured by pChIP. Utf1 protein (Cluster 1) was central to the cluster and with connections to most other proteins within cluster 1 (including Edc4). Gephi (version 0.8.1) was used to visualise the network. Bar-plots of Utf1, Edc4 and Edc3

proteins with pChIP SILAC ratios across all experiments. H/L SILAC ratios were inverted in the reverse (to be similar to forward) experiments and represented as indicated. pChIP-SILAC ratios are plotted in log scale and numbers of peptides are represented next to ratios on barplot.

6.2.7. Proteins important for network and properties

We next explored the pChIP-network to unravel and identify proteins that are inherently important for its architecture. Eigenvector centrality is one approach to identify important nodes in networks, in which a score is defined for every node in a network that measures how influential that node is for the network. A node is considered influential if it is connected to other influential nodes (One noted application of eigenvector centrality is Google PageRanks that ranks and arranges which webpages are important). The eigenvector centrality scores were measured for every cluster independently, and for the entire network.

We first looked at the proteins that had highest eigenvector centrality for the complete network and also individually for each cluster (Fig. 6.10). The proteins with highest eigenvector centrality were: Rps9, a component of mRNP granule complex known to interact with several polyA-binding proteins involved in RNA export processes; Aurk2 (Aurora kinase B), a serine/threonine kinase that plays important roles in mitosis; and Cdk7, also a serine/threonine kinase that is involved in RNAPII mediated transcription. Cdk7 is the enzyme that catalyzes the phosphorylation on serine residues (S5p and probably S7p) and directly interacts with Rpb1-CTD.

Network	Cluster 1	Cluster 2
Eef1d;mCG_22130	Pole;Pole1	Jmjd1b;Kdm3b
Rps9	Gtpbp9;Ola1	Sir2l1;Sirt1
Ark2;Aurkb	Prdm2;RP23-185C16.1-001	Ywhab;mCG_5429
Acta;Acta1	Cxn-43;Gja1	Gm16409;mCG_22088
Cdk7;Cdkn7	Utf1;mCG_21628	mCG_12245;Rbbp4

Cluster 3	Cluster 4	Cluster 5
Eef1d;mCG_22130	Polr2e;mCG_13422	H1f5;Hist1h1b
Rps9	Polr2c;Rpo2-3	Hist1h2bp;H2b-f
Ark2;Aurkb	Polr2h	Gtf2f1;mCG_5591
Acta;Acta1	Ell;mCG_23118	H1ft;H1t
Gm5451;Gm9822	Polr2b;rpB2	H1f3;Hist1h1d

Cluster 6	Cluster 7	Cluster 8
D10Wsu52e	Ptpn14;Safb2	Ddx23;mCG_18410
Hnrpq;Nsap1	Cdc2l5;Cdk13	Rbm17;Spf45
Hnrpdl;Jktbp	Np220;Zfml	Bcas2;Dam1
Tial1;mCG_21017	Hnrnpc;Hnrpc	Smu1;mCG_9820
Ptbp1;PTB4	Elavl1;Elra	Cwc15;Ed1

Figure 6.10. Identifying proteins most important for network structure. Proteins with highest eigenvector centrality, *i.e.* proteins that influence most of the network and the connections thereby being most important for network construction. Proteins with highest eigenvector centrality for the network and for individual clusters are highlighted in table.

We used another measure to find proteins most important for the connection between clusters by analyzing the sub-networks in the network model of partition. We identified bridge nodes (*i.e.* pairs of connecting nodes from two different clusters) and computed the shortest paths between every pair of nodes from the two clusters. The most important bridge nodes are those through each most shortest paths go through. Bridge edges are important because they highlight the transition between clusters. Fig. 6.11 lists all the bridge nodes and scores for all clusters. Remarkably, we see interesting proteins that bridge different clusters including Eed (Polycomb protein cluster 2 and 3), RNAPII subunits (Polr2a and Polr2b: cluster 3 and 4) and as expected components of spliceosome coordinate transition between cluster 7-8.

Protein	Score	Protein	Score
Cluster 1		Cluster 2	
Baz2a;Kiaa0314	0.46	Esg;Tle3	0.45
Cnot1;Kiaa1007	0.35	Tif1;Tif1a	0.32
Tjp1;Zo1	0.13		
Cluster 2		Cluster 3	
Eed	0.10		
Cluster 3		Cluster 4	
Tcea1;Tceat	0.25	Ell;mCG_23118	0.24
Smc2;Smc2l1	0.18	Polr2b;rpB2	0.18
Ints4	0.17	Polr2a;Rpii215	0.10
Fam129c;Bcnp1	0.11		
Cluster 3		Cluster 5	
Rpl18;RPL18	0.31	RP23-59M10.1-002;Th1l	0.22
Mat1;Mnat1	0.23	H1f5;Hist1h1b	0.16
Polr2l	0.21	Rbm12;mKIAA0765	0.14
Cluster 3		Cluster 6	
Cstf2t;Kiaa0689	0.11	D10Wsu52e	0.14
Rps6kl1	0.11	Fact140;Factp140	0.12
Kiaa1470;Rcc2	0.10		
Cluster 4		Cluster 5	
D15Ertd682e;Rpap3	0.86	RP23-59M10.1-002;Th1l	0.68
		Kiaa1111;Phf8	0.18
Cluster 4		Cluster 6	
Ell;mCG_23118	0.33	Ssrp1;RP23-232L15.5-001	0.53
Polr2b;rpB2	0.26	Kiaa0852;Morc2a	0.15
Polr2j;mCG_1879	0.12	Supt5h	0.12
		Nosip	0.10
Cluster 5		Cluster 6	
Rbm12;mKIAA0765	0.59	D10Wsu52e	0.31
Kiaa1111;Phf8	0.22	D14Abb1e;Kiaa1105	0.18
		Tox4;mKIAA0737	0.13
Cluster 6		Cluster 7	
Tardbp;Tdp43	0.10	Hnrnpa0;Hnrpa0	0.13
Cluster 6		Cluster 8	
Pc4;Rpo2tc1	0.25	Cwc27;Sdccag10	0.24
Kiaa0907	0.16	Cpsf6	0.10
Reps1	0.12		
Rbm8;Rbm8a	0.10		
Cluster 7		Cluster 8	
Snrpf;mCG_4969	0.36	Sr140	0.23
Lsm8;Naa38	0.12	Snrpd3	0.13
Cpsf5;Nudt21	0.10		

Figure 6.11. Identifying bridge nodes – Edges that link nodes between two clusters thereby maintaining network architecture. Bridge nodes that form at least 10% of shortest path between two clusters are listed in table.

6.2.8. Robustness of network analysis of pChIP datasets

6.2.8.1 Stability of network when using smaller number of pChIP datasets: only reverse or forward datasets

Our pChIP experimental setup was highly redundant and composed of replicate complimentary experiments. To test the robustness of our clustering and network results, we analysed for comparison a subset of the pChIP datasets (either using only the forward or reverse replicate experiments. We also analysed a random choice of replicates from our pChIP MS dataset and the results were similar (data not shown)

Applying the clustering to only the five reverse or the five forward datasets produced 9 or 7 clusters, respectively, with similar structure to the 8 clusters observed for the full dataset. To assess the reproducibility of our pChIP analyses, we compared the distribution of proteins in different clusters using only 5 datasets with the clusters observed for the whole pChIP dataset (Fig. 6.12). Although shuffling of proteins was detected for the reverse dataset, the major changes occurred within Clusters 5, 6 and 7 and between 2 and 3 (relative to the whole dataset). Using the five forward datasets, we observe that the core architecture is retained (Fig. 6.12), with only finer separation of Cluster 6 (seen in the Forward clustering) into two clusters (named Clusters 5 and 6 in the 9 dataset clustering). Reassuringly, the core components of the clusters were retained when using only 5 or 9 datasets, highlighting the reproducibility of our pChIP experiments and the robustness of our clustering analyses.

A

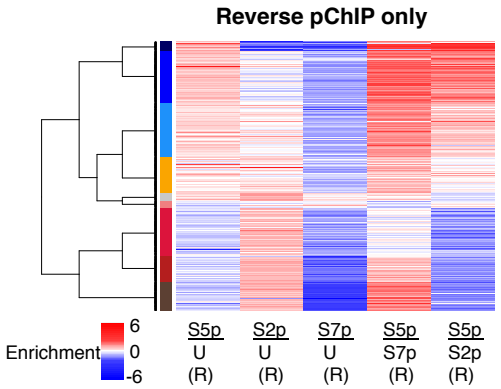
All reverse pChIP experiments

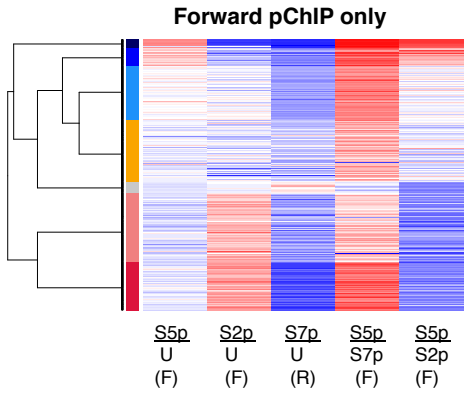
Whole dataset	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	9	5	0	0	0	0	0	0	0
C2	10	115	28	7	1	0	0	0	0
C3	3	8	102	69	2	2	3	1	2
C4	0	0	3	0	0	12	0	0	0
C5	0	0	0	3	16	2	1	0	0
C6	0	0	1	8	0	1	16	47	0
C7	0	0	0	0	0	0	10	65	1
C8	0	0	1	3	0	0	34	6	70

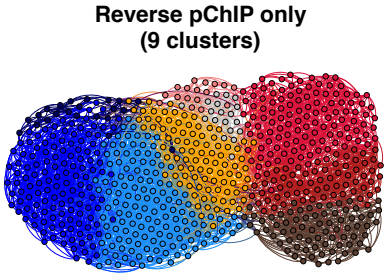
B

All forward pChIP experiments

Whole dataset	C1	C2	C3	C4	C5	C6	C7
C1	14	2	0	0	0	0	0
C2	3	39	56	41	0	8	3
C3	1	0	66	84	0	21	8
C4	0	0	0	11	3	1	0
C5	0	0	1	0	21	3	1
C6	0	0	5	7	1	53	3
C7	0	0	0	3	0	71	1
C8	1	1	0	2	0	6	100







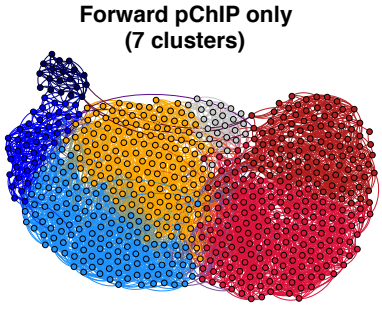


Figure 6.12. Comparing the robustness of clustering and network analysis by using a subset of pChIP datasets. Minimum number of experiments to obtain enrichments for S5p, S7p and S2p were taken and clustering/network analysis was repeated. (A and B) Clustering with whole data set (9 pChIP datasets) compared clustering with either five reverse only or forward only pChIP experiments. Protein network for both reverse only (9 clusters) and forward only (7 clusters) highlights the distribution of clusters and proteins within. Remarkably, the core architecture of clustering is retained irrespective of the replicate experiments used.

6.2.8.2 Comparing all proteins with the subset of proteins most consistently detected at least once in all pChIP pairs.

To further test the reproducibility and robustness of pChIP SILAC analysis, we repeated the clustering and network analyses using a smaller subset of 446 proteins, that were most consistently detected across the nine pChIP datasets (*i.e.* proteins identified in one of the Forward or Reverse datasets for all pChIP combinations, *i.e.* in four or five experiment pairs).

With the 446-protein dataset, we identified 7 robust and stable clusters compared to 8 clusters in whole dataset. Reassuringly, the core architecture of the clustering is retained and remarkably the core proteins between clustering are unchanged (Fig. 6.13). We observed that with the 700-protein dataset, the additional information allows for finer partitioning of the dataset and network properties (Fig. 6.13 clusters 6 and 7 – whole dataset), as expected due to the presence of replicate information.

Protein detected in four/five experiment pairs (446 proteins)

Whole dataset

	C1	C2	C3	C4	C5	C6	C7
C1	6	0	0	0	0	0	0
C2	0	85	3	0	1	0	0
C3	0	1	108	0	10	0	0
C4	0	0	0	1	10	0	0
C5	0	0	0	15	0	0	0
C6	0	0	1	0	24	20	3
C7	0	0	0	0	0	62	0
C8	0	0	0	0	0	1	94

Figure 6.13. Comparing clustering portioning using the proteins detected consistently in at least one of the pChIP experimental replicates in 4 or 5 experimental pairs to the whole dataset clustering. The core architecture of the clustering is retained between both datasets using 445 proteins most consistently detected across pChIP experiments. Cluster 1 is still observed, albeit with 6 proteins, all of which are conserved between clustering's. Very few proteins are detected in cluster 4 of the whole dataset. For the proteins detected they mostly reshuffle between adjacent datasets; additionally with whole dataset there is a finer separation of proteins (cluster 6 and 7 – whole dataset).

We next investigated the network architecture and connections within the pChIP-network. On visual inspection, the networks look entirely similar to each other and their partitions (Fig. 6.14). In the whole dataset, we had finer partitioning and additional proteins (254 proteins) spread non-preferentially across all clusters. Notably, we identify a large S5p cluster (cluster 3 whole dataset that is partitioning of cluster 2 (446 proteins) and additional proteins (254 proteins). Looking at the average intensity profiles for whole dataset and 446 proteins, we observed similar profiles across all experiments. The gradient separation of proteins again is best described by S5p/S2p experiment in both the datasets.

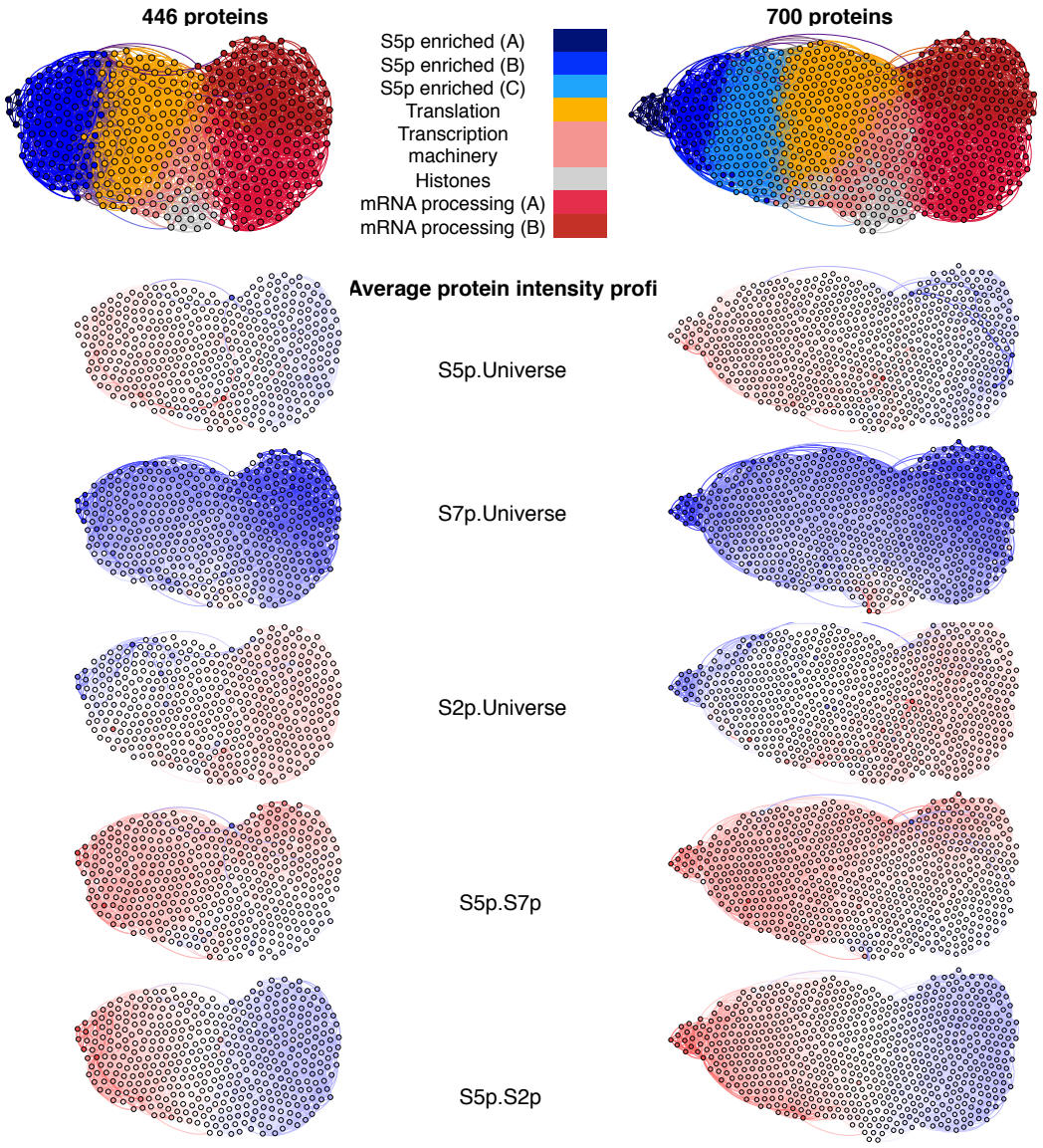


Figure 6.14. Comparing protein-network between whole dataset (700 proteins) and most consistently detected proteins in at least one of the pChIP pairs (446 proteins). 445 proteins were found most consistently detected in all experiments (5 or 4 pairs) and partitioned into 7 clusters. Comparing the different networks, the overall shape and architecture between the two clustering's is maintained; additionally, the similarity of average intensity plots highlights the robustness of our pChIP method.

6.3. Discussion

Our analyses of the proteome associated with chromatin-bound proteome in association with RNAPII complexes modified on S5p, S7p or S2p show that RNAPII is involved in a diverse range of biological process fundamental for mES cells. Using novel clustering and network algorithms, we have further

detected patterns and unravelled associations not previously identified. The novelty of our clustering and network algorithm is that all analysis is done on pChIP-SILAC ratios in an unbiased manner without the need for any prior information on proteins (including protein length, subunit structure, size, abundance or properties); this approach requires redundancy (replicates) and slight differences between experiments (e.g. pairwise and universe approaches).

6.3.1. Clustering sensitively detects a gradient partitioning of protein association with chromatin bound by different RNAPII variants

From our simple pChIP classification system, we observed proteins enriched specifically for S5p only and proteins detected with a combination of modifications. Our clustering and network algorithm sensitively detects the association of proteins relative to RNAPII and partitions them based of their pChIP-SILAC ratio. The gradient separation flows from proteins enriched for S5p only (no S7p or S2p) to proteins with basal RNAPII modifications and finally proteins enriched for S5p and S2p. This remarkable gradient separation and novel partitioning was not anticipated and highlights the sensitivity and need for our systems approach and contrasts from conventional methods. Proteins with varying ratios across experiments which were difficult to understand in the simple classification used in Chapter 5 (example Paf1 complex in Fig. 5.17; Cluster 5-pink colour in Fig. 6.5) could be elucidated by our clustering approach as robustly detected in correlation with the behaviour of Rpb subunits without any specific enrichment for RNAPII modifications.

6.3.2. Robust partitioning unravels novel patterns within S5p proteins and common proteins.

The clustering algorithm dissects the wealth of information contained in pChIP-SILAC ratios and identifies novel patterns that are biologically meaningful including stoichiometry of interactions and relative affinity to

RNAPII. Moreover we are able to demarcate protein association with RNAPII on chromatin based on levels of enrichment of pChIP-SILAC ratios.

We identify two clusters (Cluster 7 and 8) both enriched for S5p&S2p with varying levels and consisting of co-transcriptional machinery (including RNA processing, splicing, polyadenylation and export). Interestingly, we observe enrichment for export factors in Cluster 8 than Cluster 7 suggesting the transition of transcriptional process. We also distinctly separate RNAPII subunits (cluster 5) from histones (cluster 4) even with basal levels of RNAPII modifications. We demarcate and robustly identify three clusters associated with RNAPII-S5p only based on their affinities of interaction. Both Polycomb and replication proteins are identified in cluster 2, however cluster 3 consists of groups of proteins involved in other repressive complex, chromatin remodellers and metabolic proteins. It is apparent that clustering senses the variation of proteins association between cluster 2 and cluster 3. Remarkably, we also identify a small robust cohort of proteins (cluster 1) with strongest enrichment for RNAPII-S5p including an essential stem cell transcription factor, chromatin remodellers and proteins of uncharacterised functions previously unknown to functionally interact with RNAPII. We suspect that these proteins (given their roles in fundamental processes and S5p-only association) regulate important stem cell processes and such systems approach is required to uncover them.

6.3.3. Systems approach uncovers novel biological insights

The clustering and network algorithm apart from partitioning the proteins also provides high level of information including the nodes important for cluster architecture and within each cluster. The edge weight further signifies the similarity of association relative to RNAPII and further predicts co-association along with RNAPII modification. These parameters play an important role for choosing candidates and understanding their regulation. For example, we observe Ring1b protein directly connected with Ogt protein in cluster 2 with a

stringent edge weight (Fig. 6.8). Not surprisingly, very recently Ogt has been shown to interact with Polycomb proteins and also RNAPII-CTD (Myers *et al.* 2011; Ranuncolo *et al.* 2012). Other linked proteins in cluster 2 would provide important candidates that might be involved during the Ogt-Polycomb-RNAPII interplay.

Our large pChIP experiment was designed to have comprehensive and complimentary replicate information encoded within it. Performing analysis with replicates (*i.e.* forward only or reverse only or random selection), we robustly observe the retention of core network architecture and highlight weaker associations. In summary, the analyses presented in this chapter clearly demonstrate the robustness, strength and sensitivity of our systems biology approach to capture and dissect RNAPII-bound chromatin interactions.

7. Comparing RNAPII pChIP with mRNA bound proteome dataset and mitotic RNAPII

7.1. Research motivation

Proteome-ChIP captures cohorts of interactions occurring on chromatin. In Chapters 5 and 6, I describe the experimental strategy, simple classification analysis and systems biology approach to dissect and unravel these interactions and their biological relevance. My aim in this chapter was to further explore the basis of these biological interactions by distinguishing mRNA-protein interactions from protein-protein interactions and other interactions. I also aimed to integrate the types of interactions to our system biology analysis using both simple classification and network analysis. Borislav Vangelov and Prof. Mauricio Barahona performed the network analysis (Figs. 7.5 and 7.8) in an active collaboration between our laboratories.

7.2. RNAPII regulation on chromatin

RNAPII transcription is a highly regulated and dynamic process involving several components that sense the local environment and cascade a range of feedback, feed-forward and downstream processes.
A plethora of dynamic associations with RNAPII or chromatin of a range of regulatory complexes occurs during transcription; these associations include DNA-protein interactions on chromatin, protein-protein interactions (forming protein complexes) and protein-RNA interactions (e.g. direct binding of the RNA processing machinery to the nascent RNA). During transcription, the Rpb1-CTD is known to act as a scaffold for interactions (protein-protein) that further regulate distinct stages of transcription and interactions with DNA (chromatin remodellers), mRNA (splicing factors) and protein (RNAPII-recycling).

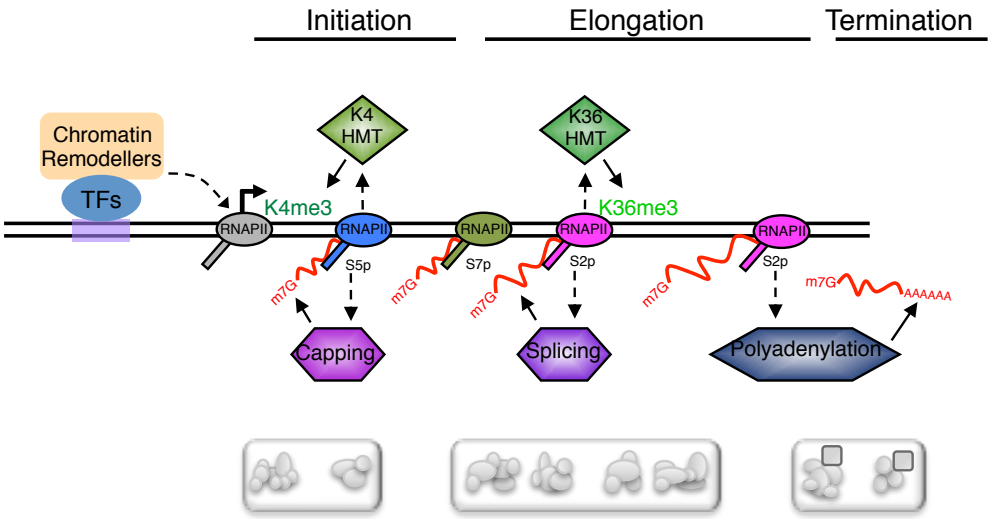


Figure 7.1. Schematic representation of steps involved during transcription and different types of interactions. Transcription involves range of interactions occurring on proteins, DNA and mRNA. During transcription, distinct modifications on Rpb1-CTD assist in recruitment of range of chromatin modifiers, capping enzyme, RNA processing factors, splicing machinery and transcription termination machinery to produce a stable mRNA transcript completing one round of transcription.

7.3. Capturing different types of interactions.

With the advent of newer technologies and significant improvements in high-throughput techniques like mass spectrometry and next generation sequencing, newer methods are developed to identify different interactions and their biological validity. DNA-Protein interactions were conventionally detected by electrophoretic mobility shift assay (EMSA; (Garner and Revzin 1981), DNase footprinting (Galas and Schmitz 1978) or yeast One-hybrid assays (Ouwerkerk and Meijer 2001). More recent, high-throughput genome-wide methods, like DNA-ChIP, sensitively identify and provide protein occupancy maps across the whole genome (Gilmour and Lis 1984; Solomon *et al.* 1988). Methods for detection of RNA-protein associations include more recent high-throughput technologies including RNA-immunoprecipitation, PAR-CLIP and further developments like HITS-CLIP and CLIP-Seq (Keene *et al.* 2006; Hafner *et al.* 2010; Konig *et al.* 2011). Conventional and more recent large-scale proteomic studies have also identified and elucidated protein-

protein interaction maps leading to an availability of resourceful databases (Phizicky and Fields 1995; Prieto and De Las Rivas 2006).

7.4. Results

The protein interaction landscape captured by pChIP using RNAPII antibodies includes protein-protein, protein-DNA, protein-RNA and other interactions dependent on chromatin state. I next asked if we could further dissect these interactions and highlight them. The obvious approach is to perform the comprehensive pChIP experiment with RNase to identify the non-RNA bound interactions, however with caveats including overall lower number of proteins (per sample volume), normalization issues and variance in stoichiometry of proteins and their pChIP-SILAC ratios; even then, treatment with RNAse will not remove proteins that interact both with RNA and with protein. To identify the proteins identified in our network that also directly associate with RNA, we instead compared our dataset with a published human mRNA-bound proteome (MBP) dataset produced using (APPROACH and what it captures) (Baltz *et al.* 2012).

To capture and highlight interactions with RNAPII in our pChIP dataset, we compared our pChIP dataset with a dataset of RNAPII interactors produced in native chromatin-depleted protein extracts. Our lab has previously published the protein interactome of RNAPII complexes isolated in native conditions during mitosis in HeLa cells, a transformed human cell line (Moller *et al.* 2012), after separation from mitotic chromosomes and DNAse I treatment. Briefly, the majority of the interactions captured in this dataset should represent protein-protein interaction (non-chromatin).

7.4.1. Summary of steps involved in comparing mouse RNAPII pChIP dataset with published human datasets

To compare the published human datasets with pChIP dataset from mES cells, I followed a simple strategy (Fig. 7.2). Published MS protein datasets from PAR-CLIP or mitotic RNAPII co-immunoprecipitation were converted from the human protein/gene identifiers to corresponding orthologs, *i.e.* protein/gene identifiers in mouse. The orthologs in the two published datasets were then compared with simple pChIP classification and also overlaid on pChIP-network to visualise their distribution.

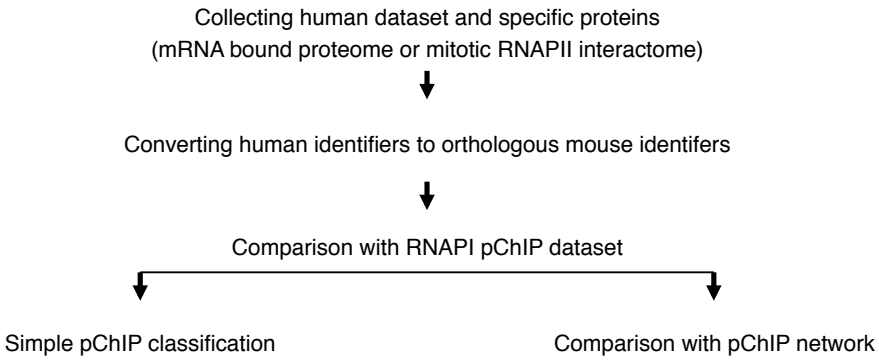


Figure 7.2. Steps involved in comparing human published datasets with RNAPII pChIP dataset. Proteins identified in published human dataset (mRNA bound proteome and Mitotic RNAPII proteins) are first converted to orthologous mouse identifiers and compared with simple pChIP classification (Fig, 5.12) and with pChIP network (Fig. 6.7)

7.4.2. Comparing human mRNA bound proteome (MBP) dataset with RNAPII pChIP dataset.

Conversion of the original human MBP dataset into ortholog protein identifiers in mouse gave rise to a list of 760 proteins. Interestingly, we observed a significant overlap of 40% (301 proteins) between MBP and the 702 proteins identified by pChIP (Fig.7.2), suggesting that their strategy may capture many RNA processing and transport regulators that associate with RNA on chromatin during transcription. Looking at GO terms reassuringly, the

common proteins were all enriched in RNA-associated terms including ‘mRNA processing’, ‘RNA splicing’ and ‘mRNA polyadenylation’. Interestingly, ribosomal proteins were also common and enriched GO terms included ‘translation’ and ‘ribosome biogenesis’, which may be explained in this case due to the association of mRNA with the functional ribosome.

Looking at the remaining pChIP-only proteins (436 proteins), we observed significant enrichment for DNA-associated processes including ‘chromatin assembly’, ‘histone mono-ubiquitination’, ‘histone acetylation’ and ‘DNA replication’ (Fig. 7.3).

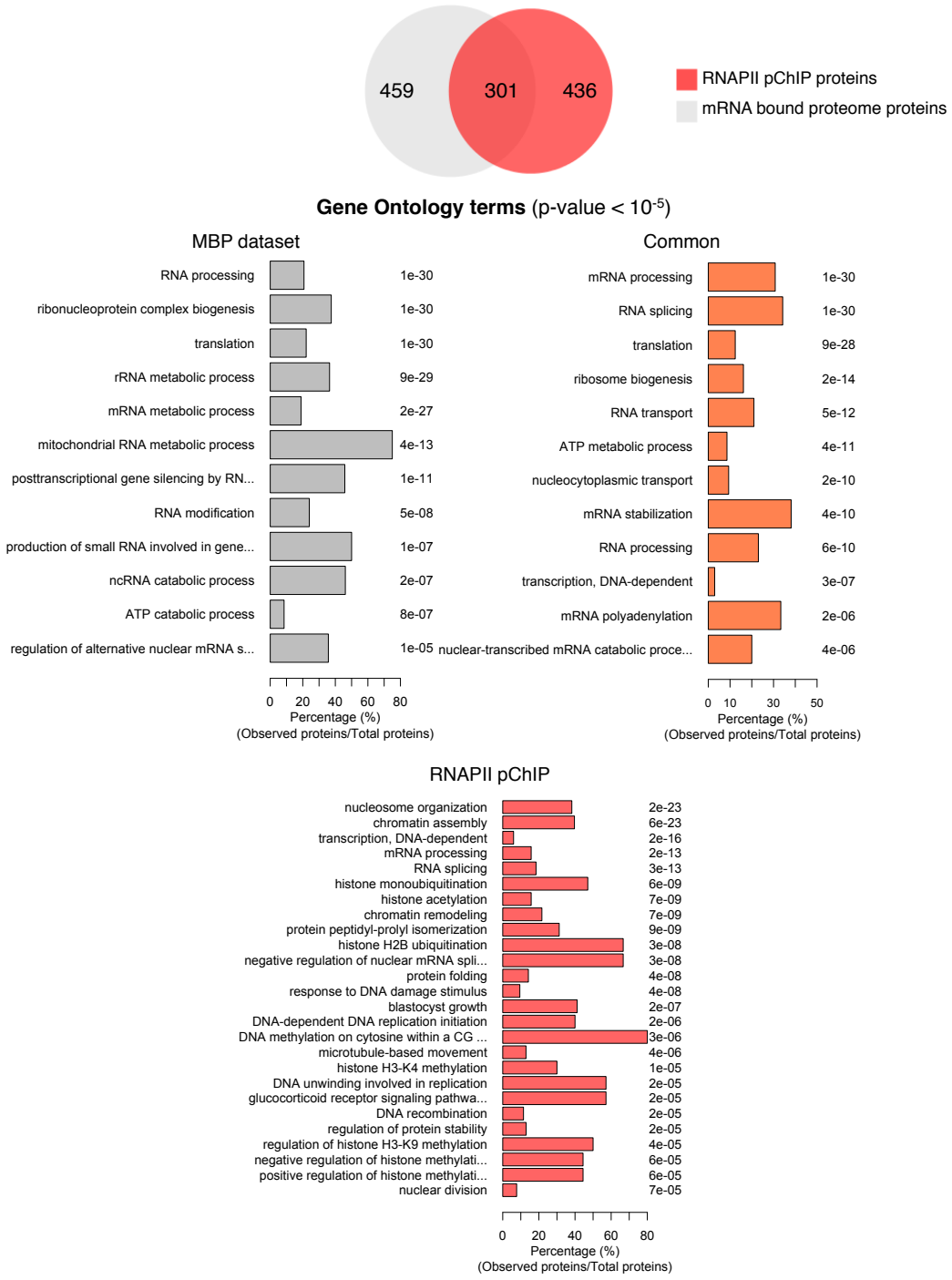
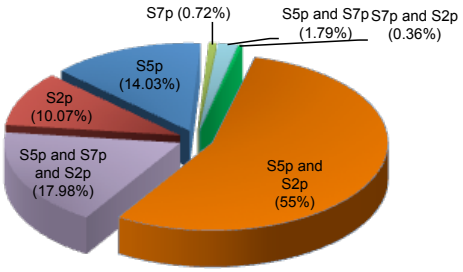


Figure 7.3. Comparing of proteins identified in mRNA bound proteome (MBP) dataset and RNAPII pChIP. Venn diagram highlights significant overlap of proteins (301 proteins) between MBP dataset and RNAPII pChIP. Significant GO terms for proteins enriched in MBP dataset, common and RNAPII pChIP proteins are represented by barplot. Grey, orange and red barplots represent GO terms for MBP proteins, common and RNAPII pChIP proteins respectively.

7.4.3. Comparing MBP dataset with simple pChIP classification

Successful rounds of transcription involve phosphorylation of RNAPII (*i.e.* S5p followed by S7p and S2p) on chromatin thereby recruiting co-transcriptional machinery to stabilize and process the nascent RNA to produce mRNA. Therefore, we anticipated that most of the interactions common to pChIP and MBP datasets would be associated with S2p marks (*i.e.* 'S5p & S7p & S2p', 'S5p & S2p' and 'S2p only'). When compared the proteins common to MBP and pChIP (301 proteins; Fig. 7.3) to the simple pChIP classification shown in Chapter 5 (Fig. 5. 12), we find as anticipated that the majority of MBP proteins were enriched with 'S5p & S2p' (55% of 301 common proteins) and 'S5p & S7p & S2p' (18%). We plotted a pie chart of the distribution of proteins (Fig. 7.4) and the table highlights the number of proteins. We observe clear enrichment for transcriptional elongation (S2p processes) with the MBP dataset and this further highlights the specificity of pChIP and its multiple applications to identify and dissect interactions.



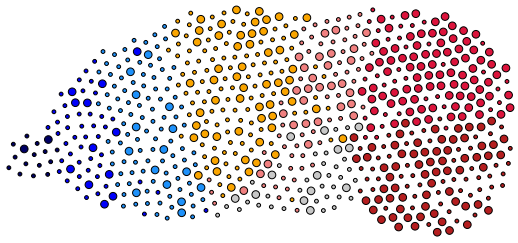
Classification	Total number of proteins	Number of proteins found in MBP dataset
S5p	185	39 (14%)
S7p	15	2 (1%)
S2p	28	28 (10%)
S5p and S7p	5	5 (2%)
S5p and S2p	156	156 (55%)
S7p and S2p	4	1 (0%)
S5p and S7p and S2p	92	52 (18%)

Figure 7.4. Preferential association of MBP proteins to elongating RNAPII (S5p&S2p) visualised by pie chart. Pie chart represents different RNAPII classifications and percentage of proteins in each group. Table consisting of number of protein identified in MBP and their percentages along with number of proteins in each pChIP classification.

7.4.4. Overlaying common proteins pChIP protein network

We next asked what was the position of the MBPs common to our pChIP dataset on the protein community network presented in Chapter 6 (Fig. 6.5), in

particular to asked if the proteins were preferentially associated with 'S5p & S2p' clusters. Overlaying the common proteins on the pChIP network, we observed that proteins were specifically enriched in clusters 4, cluster 7 and cluster 8 (Fig. 7.5). Cluster 7 contained 85 proteins (31%), while cluster 8 and 4 contained 66 (24%) and 56 (20%) proteins. The preference for MBP proteins for 'S5p & S2p' clusters is visually striking (enlarged nodes in Fig. 7.5) and further highlights the specificity and necessity to use machine-learning approach to unravel interaction dependencies.



pChIP clusters	Total number of proteins	Number of proteins found in MBP dataset
S5p enriched (A)	16	2 (1%)
S5p enriched (B)	57	12 (4%)
S5p enriched (C)	109	15 (5%)
Translation	154	56 (20%)
Transcription machinery	83	29 (11%)
Histones	44	13 (4%)
mRNA processing (A)	119	85 (31%)
mRNA processing (B)	118	66 (24%)

Figure 7.5. MBP are preferentially located in clusters with S5p&S2p. MBP protein overlaid on pChIP network with enlarged nodes representing proteins identified in both dataset. Preferential association is observed in clusters (cluster 7,8 and 4) containing S5p&S2p. Table consisting of number of proteins in each category.

7.4.5. Comparing human RNAPII-mitotic interactome with RNAPII pChIP dataset.

We next compared our pChIP dataset with RNAPII-mitotic interactome dataset (Moller *et al.* 2012) to identify and highlight protein-protein interactions from other interactions. We identified 409 mouse protein orthologs in the Moller dataset, and observed an overlap of 200 proteins between RNAPII-mitotic interactome and RNAPII pChIP. It is worthwhile remembering that mitotic RNAPII is heavily phosphorylated on S2p, in spite of being dissociated from chromatin, a process thought to prevent reinitiation of RNAPII complexes during mitosis; in agreement with this RNAPII state, Moller and colleagues showed that their dataset is enriched for proteins involved in mRNA processing and suggested that the association in mitosis between RNAPII and the RNA processing machinery, in the absence of active transcription,

could be important for the coordinated import of nuclear components once the nuclear membrane reforms and to promote efficient activation of mRNA synthesis in G1 phase of the cell cycle.

Looking at GO terms for common proteins, we enrich for protein-protein interaction with biological functions in ‘mRNA processing’, ‘translation’ and ‘ATP metabolic processes’ (Fig. 7.6). The pChIP-only proteins as expected are now enriched for a range of RNA, chromatin and nuclear processes including ‘mRNA processing’, ‘chromatin assembly’ and ‘DNA replication’ (Fig. 7.6).

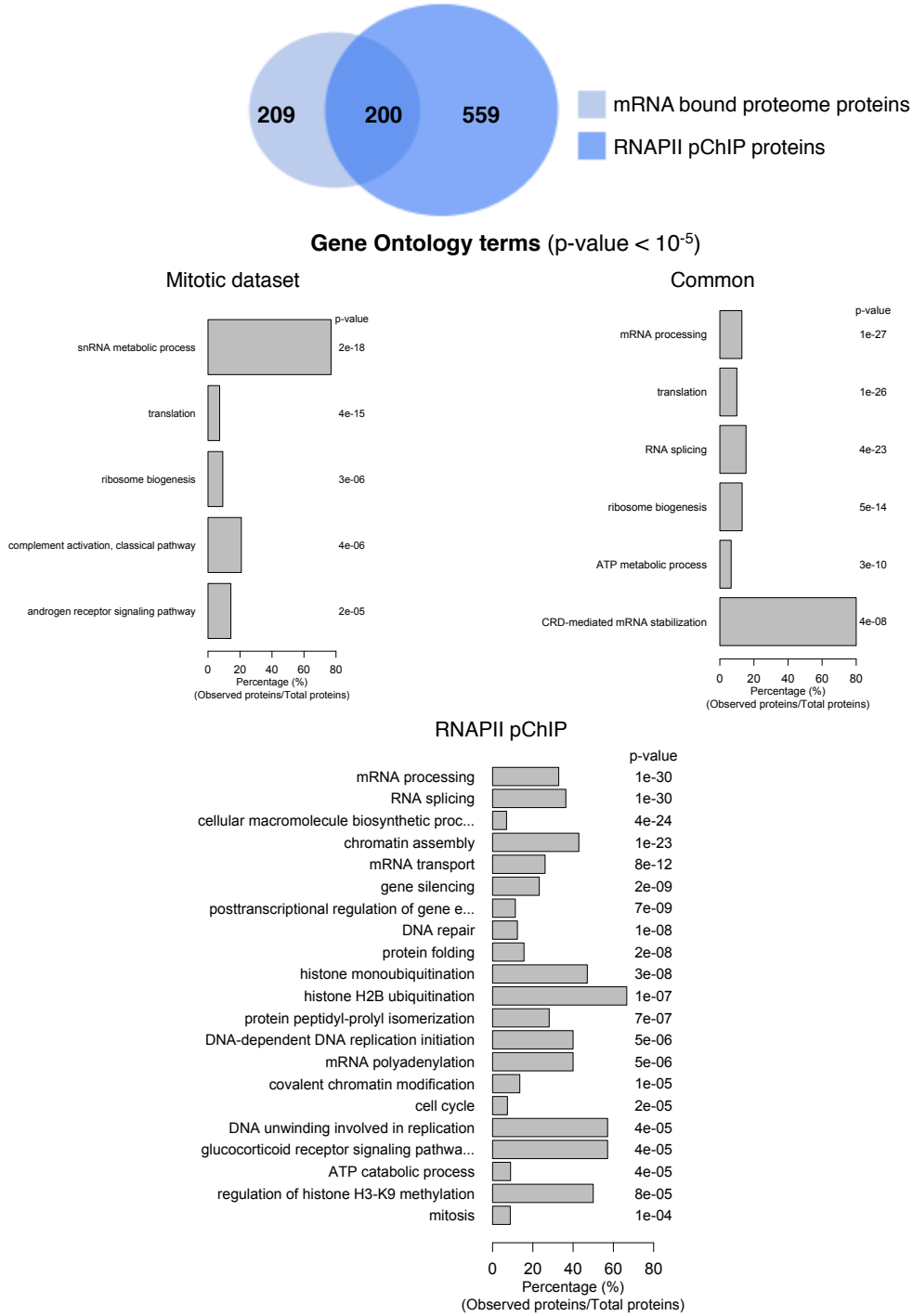
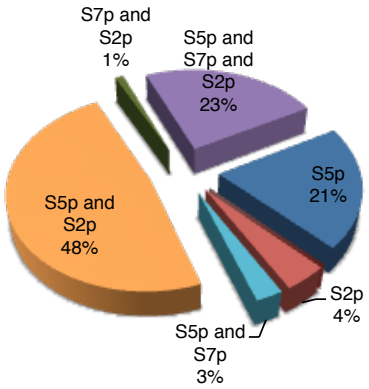


Figure 7.6. Comparing the mitotic RNAPII proteome with RNAPII pChIP dataset. Venn diagram with proteins identified in both RNAPII pChIP dataset and mitotic RNAPII interactome. 200 proteins were identified in both datasets. Significant GO terms for proteins enriched in different categories are represented as barplots.

7.4.6. Comparing mitotic RNAPII interactome with simple pChIP classification and pChIP-network.

Unlike RNA-protein interactions and comparison with MBP dataset, protein-protein interactions can occur on DNA during different stages of transcription. Consistently, we observe enrichment in different pChIP classifications, 48% of the proteins associated with RNAPII in mitosis were enriched with 'S5p & S2p', 23% proteins were enriched for 'S5p & S7p & S2p' and 21% proteins were enriched for S5p (Fig. 7.7). However, it is interesting to observe that many more proteins in the GO term 'mRNA processing' are identified in the RNAPII pChIP dataset (35% of 200 proteins common to both dataset), than in the mitotic RNAPII proteome dataset (15% of 200 proteins common to both dataset), suggesting that although some processing components may remain associated with RNAPII during mitosis, that the majority only associates co-transcriptionally. This highlights the power of pChIP in freezing and retaining co-associations that only exist on chromatin, and are likely missed after protein purifications. Comparisons between the mRNA processing proteins present throughout mitosis, and those only present on chromatin, may yield interesting insights on the dynamics of recruitment of the RNA processing machinery, but were not further explored in this thesis.

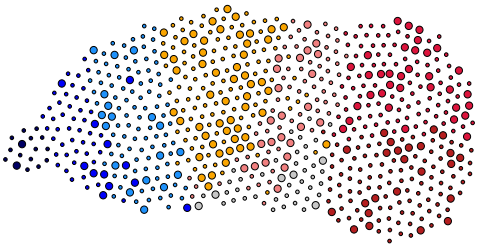


Classification	Total number of proteins	Number of proteins found in Mitotic RNAPII dataset
S5p	185	38 (21%)
S7p	15	0 (0%)
S2p	28	8 (4%)
S5p and S7p	5	5 (3%)
S5p and S2p	156	89(48%)
S7p and S2p	4	2 (1%)
S5p and S7p and S2p	92	43 (23%)

Figure 7.7. Subsets of mitotic RNAPII interactions are captured by RNAPII pChIP. Pie chart of proteins identified in both mitotic RNAPII proteins and RNAPII pChIP. Colours represent different pChIP classification with percentage of proteins

identified in mitotic dataset. Table on right consists of protein numbers in pChIP classification and shared mitotic proteins.

We next overlaid the RNAPII-mitotic proteins onto pChIP-network to observe whether there is any preferential association of the RNAPII-bound proteins in mitosis with the pChIP classification (Fig. 7.7). We observed random distribution of common proteins (enlarged nodes in Fig. 7.8) spread across different clusters, which suggests that proteins across various processes retain association with RNAPII through mitosis. Future work may highlight whether these proteins are pioneers in establishing the chromatin-associated proteome identified with RNAPII, or whether they simply result of our technical ability to detect some proteins and not others.



Classification	Total number of proteins	Number of proteins found in mitotic RNAPII dataset
S5p enriched (A)	16	2 (12%)
S5p enriched (B)	57	11 (19%)
S5p enriched (C)	109	17 (15%)
Translation	154	53 (34%)
Transcription machinery	83	23 (28%)
Histones	44	4 (9%)
mRNA processing (A)	119	25 (21%)
mRNA processing (B)	118	34 (8%)

Figure 7.8. Overlaying mitotic RNAPII proteins over pChIP network. Proteins common in both mitotic RNAPII and pChIP are overlaid and represented by enlarged nodes in pChIP network. No specific enrichment of mitotic proteins is observed in any clusters.

The main advantage of comparing pChIP dataset with published datasets is that we can dissect and predict the interaction landscape at a minimum cost of bioinformatics analysis and no experimental cost. The pChIP network offers a visual representation of the interactions and comparison between different networks (Figs. 7.5 and 7.8), helping to find DNA-protein interactions and interactions dependent on chromatin state.

7.5. Discussion

RNAPII regulation in mES cells is complex, dynamic and not only restricted to actively transcribing genes but also to PRC-repressed genes and other fundamental biological processes as unravelled by pChIP. In Chapters 5 and 6, I have demonstrated that pChIP captures and dissects the biological basis of RNAPII chromatin-bound proteome and unravels its dependencies. Comparisons of the pChIP datasets with the MBP and mitotic-RNAPII interactome datasets shed further insights into the types of interactions represented in the pChIP classification and network. These comparisons also allow us to correlate and predict which interactions (DNA-protein, RNA-protein or protein-protein) serve as primary steps and further cascade of downstream interactions (secondary). For example, chromatin remodellers involved in transcription elongation (maintaining open chromatin) are not identified in MBP or mitotic RNAPII datasets, however many, but not all, splicing factors are identified in MBP and mitotic-RNAPII dataset (Interactions with RNA and proteins).

A comparison of pChIP with published datasets provides a cost effective option of asking specific questions and obtaining biologically relevant insights. For example, RNA-associated proteins are enriched in S5p&S2p clusters (clusters 7, 8 and 4) and consist of splicing factors and interestingly proteins involved in translation. Intriguingly, the few proteins identified in S5p clusters (clusters 1, 2 and 3) raise interesting hypothesis including the role in abortive transcription, ncRNA regulation and also miRNA-mediated regulation.

These comparisons also provide interesting candidates that form a bridge between RNAPII-CTD, RNA and co-transcriptional machinery. And add an additional layer of information about which subunits of complexes may directly interact with RNAPII, chromatin or RNA, and the ancillary subunits. Indeed, these analyses provide us a starting point to perform further experiments including treatment of pChIP samples with RNase, removing DNA and RNA to

identify the protein only cohorts. Identification of types of interactions also allows us to robustly select candidates for further analysis and appropriate experimental design. For example, a protein-protein interaction occurring on chromatin and dependent on chromatin state is unlikely to be identified from immunoprecipitation from whole cell extract preparations or nuclear extract preparations, the standard approach currently in the literature to identify protein co-associations.

Ideally, we would like to perform RNAPII pChIP experiments that interfere with the cohorts of associated proteins, such as inhibiting of transcription elongation (with drugs DRB or flavopiridol) and global RNAPII transcription (α-amanitin). But we also believe that our comparisons with MBP dataset and mitotic RNAPII proteome provide insights into understanding and interpreting the drug inhibition results.

8. Extending pChIP using Native chromatin and Gradient pChIP for crosslinked chromatin

8.1. Research motivation

In the previous chapters, I investigated the RNAPII-bound chromatin proteome using pChIP on crosslinked chromatin, unravelling and dissecting dependencies to RNAPII. My aim in this chapter was to extend pChIP and explore the RNAPII-bound chromatin proteome in non-crosslinked native chromatin preparations and in fractionated-chromatin preparations (both lower complexity samples). I aimed to perform DNA-ChIP and pChIP on histone modifications, Polycomb proteins and RNAPII modifications on native chromatin. To unravel the chromatin proteins that co-exist with RNAPII and Polycomb at PRC-repressed genes towards investigating the RNAPII-Polycomb interplay on chromatin, I aimed to optimise the Gradient-pChIP protocol, which focuses on Polycomb proteins associated with chromatin and excludes chromatin-free Polycomb complexes.

All of the MS experiments were carried out in collaboration with Dr. Bram Snijders at Proteomics facility at MRC-CSC. Bram also helped with sample pre-processing for MS and performed all MS run time operations. Dr. Carmelo Ferrai from our group helped with guidance and assistance in gradient fractionation of crosslinked chromatin. MS analyses on gradient samples are current being carried out by Dr. Guido Mastrobuoni and Dr. Stefan Kempa at Max Delbrück Centre for Molecular Medicine (Berlin) in an active collaboration between the two laboratories.

8.2. Native chromatin

Native chromatin is prepared by micrococcal nuclease (MNase) digestion of cell nuclei that leads to chromatin resolution of nucleosomes (Hebbes *et al.* 1988). MNase cuts DNA between the nucleosomes where it is not protected by histones, proteins and multi-protein complexes; therefore, it can be

optimised to obtain nucleosomal resolution. Owing to its resolution and native extraction conditions, native-ChIP is thought to be applicable for histones, histone modifications and proteins that remain tightly associated to chromatin during the fractionation procedure. At active gene promoters, RNAPII is thought to bind at the nucleosome free region (NFR) between spaced nucleosomes and recruit pre-initiation complex (PIC). During native chromatin preparation, these interactions are thought to be lost owing to excessive purification of mono-nucleosomes.

Bivalent domains present at the chromatin of PRC-repressed genes have been identified by native-ChIP (Bernstein *et al.* 2006; Mikkelsen *et al.* 2007; Ku *et al.* 2008; Brookes *et al.* 2012). These genes are simultaneously marked by active H3K4me3 and repressive chromatin marks (H3K27me3 and H2Aub1). RNAPII at these genes assessed by crosslinked DNA-ChIP was observed in a novel S5p-only state without productive mRNA production.

In Chapter 3, the large diversity of native chromatin proteins has been demonstrated (Fig. 3.4), owing to more sensitive mass spectrometry techniques. Therefore to push the envelope, I further tested and performed DNA-ChIP and pChIP experiments on histone modifications, Polycomb proteins and RNAPII modifications (hallmarks of PRC-repressed state) using native chromatin.

8.3. Chromatin fractionation by salt gradient

Early DNA-ChIP protocols involved *in-vivo* fixation of living cells by formaldehyde, followed by purification of nuclei, sonication to break chromatin fragments of appropriate size range and purification of chromatin by cesium chloride gradient centrifugation. The chromatin fragments containing protein of interest are enriched by immunoprecipitation and finally DNA enrichment is measured after reverse crosslinking (Solomon *et al.* 1988; Orlando *et al.* 1997). The initial studies had observed that bulk crosslinked chromatin

fragments were enriched in defined density gradients, distinctly different from the density of free DNA. In addition estimation of protein-to-DNA ratios by radioactive labelled amino acids and nucleotides in bulk crosslinked chromatin confirmed the same. These studies highlighted the use of density centrifugation to purify crosslinked chromatin (nucleo-histone fractions) from free DNA and protein for DNA-ChIP.

In more recent DNA-ChIP protocols, the additional gradient separation step was removed primarily to save time and sample, without much effect on DNA yields (Schwartz *et al.* 2005). More importantly, it was reported that certain specific chromatin regions had lower density than bulk chromatin in specific cell type and these regions were not effectively captured in the bulk chromatin by gradient centrifugation (Ip *et al.* 1988; Reneker and Brotherton 1991; Schwartz *et al.* 2005). Salt fractionation has been more recently used to obtain classical transcriptionally active chromatin prepared from native micrococcal nuclease digestion of chromatin (Henikoff *et al.* 2009). Separation of nucleo-histone fractions having lower complexity (than unfractionated crosslinked chromatin) would enable pChIP to more specifically capture different chromatin states. Therefore, I aimed to first test the fractionation and perform quality control experiments to avoid any gradient biases, further performing pChIP with RNAPII and Polycomb proteins on nucleo-histone fractions to specifically identify co-existing protein cohorts.

8.4. Results

The resolution of native chromatin ranges primarily from mono- and di-nucleosomes to few nucleosomal repeats. This resolution in turn highlights the applicability of native ChIP for identifying closely associated, strong, more robust interactions occurring within the span for few hundred base pairs. The other advantage is the lack of crosslinking and therefore avoids sample (and peptide) loss during reverse crosslinking and elution steps. However, the resolution also restricts the overall amount of protein and its complexity, thereby narrowing native ChIP towards asking more specific questions. Generally, native chromatin is used to observe occupancy of histones, histone modifications and abundant transcription factors across a genome.

8.4.1. Diversity and composition of Native chromatin proteins analysed using mass spectrometry

To understand the diversity and composition of proteins in native chromatin, we first analysed input native chromatin proteins by MS. Sample preparation for MS and its analysis was done by Bram Snijders (Proteomics facility at MRC-CSC). I performed quality control experiments for the native chromatin (Coomassie staining and Agarose gel) and these results were consistent with previous results described in Chapter 3.

We identified 2070 proteins from native input chromatin prepared from ES cells, and subsequently explored the major classes of proteins observed. Briefly, we observed abundance of few groups of proteins involved in ‘Transcriptional regulation’, ‘Chromatin remodelling’, ‘Pluripotency’, ‘Metabolism’ and several groups of enzymes. Examples of proteins and their functions also listed in Fig. 8.1.

Native chromatin (2070 proteins)

Transcription components

RNAPII subunits, Splicing machinery, HnRNP's, SnRNP's Sub1,Abt1, Rprd1a, Rpap3, Cpsf6, Cstf2t, Supt16h, Paf1, Gtf3c3, Naf1

Chromatin remodellers and pluripotency

Oct4, Sox2, Dppa2, Dppa4, Drg2, Eed, Rnf2, Chaf1b, Rcor2, Suds3, Gatad2

Metabolic proteins

Hadh, Idh2, Tdh, Mdh2, Mthfd1, Aklbh5, Pdhb

Interesting proteins

Eri1 (3'-5' exonuclease), Cbx3, Cbx5

Histones

Figure 8.1. Composition of proteins identified in native input chromatin and their functions. MS analysis of native chromatin input identified 2070 proteins with functions in diverse biological processes including transcription, co-transcriptional processing, metabolism and other cellular processes. A large number of enzymes were also identified as listed.

8.4.2. Distribution of histone modifications captured on Native chromatin.

From our MS analysis of native chromatin, we observed an abundance of histones, repressive protein complexes and transcription related GO terms. In addition, we have previously identified the novel interplay between repressive histone modifications co-existing along with Polycomb and RNAPII in mES cells (Brookes *et al.* 2012). Therefore, we next decided to check the occupancy of a few histone modifications, Polycomb proteins and RNAPII modifications in our mES cells using native chromatin. Emily Brookes (from our lab) had optimised DNA-ChIP for histone modifications H3K27me3, H2Aub and H3K36me3 and these DNA-ChIP results were used as standard for positive controls while performing Polycomb and RNAPII DNA-ChIP on native chromatin.

8.4.3. H3K27me3 and H2Aub1

I first started by reproducing previous DNA-ChIP analysis of repressive histone modifications (H3K27me3 and H2Aub) across a panel of genes in

mES cells (Stock *et al.* 2007b; Brookes *et al.* 2012). H3K27me3 is a repressive histone mark, catalysed by Polycomb protein Ezh2 (PRC2 subunit) and is mainly observed as islands demarcating promoters of PRC-repressed genes and sometimes in gene body (albeit to a much lower extent). H2Aub is also a repressive histone mark, catalysed by Polycomb protein Ring1b (PRC1 subunit) and occurs to a lower extent than H3K27me3 and is enriched at promoters of PRC-repressed genes. From our lab's work, we now know that H3K27me3 is observed in 30% of gene promoters and H2Aub is enriched at 20% of gene promoters (Brookes *et al.* 2012).

I performed DNA-ChIP using highly specific antibodies directed against H3K27me3 and H2Aub and consistent with published data (Stock *et al.* 2007b; Brookes *et al.* 2012). H3K27me3 was enriched at promoters and coding (+2kb downstream of TSS) regions of only PRC-repressed genes consistent with a role as a repressive mark (Fig. 8.2). H2Aub also has similar pattern and is enriched at promoters and coding regions of PRC-repressed genes (Brookes and Pombo 2009b). No significant enrichment was observed at either active or silent genes (Fig. 8.2). Mock ChIP demonstrates the specificity of our ChIP protocol, with no detectable enrichment compared to ChIP with H3K27me3 or H2Aub.

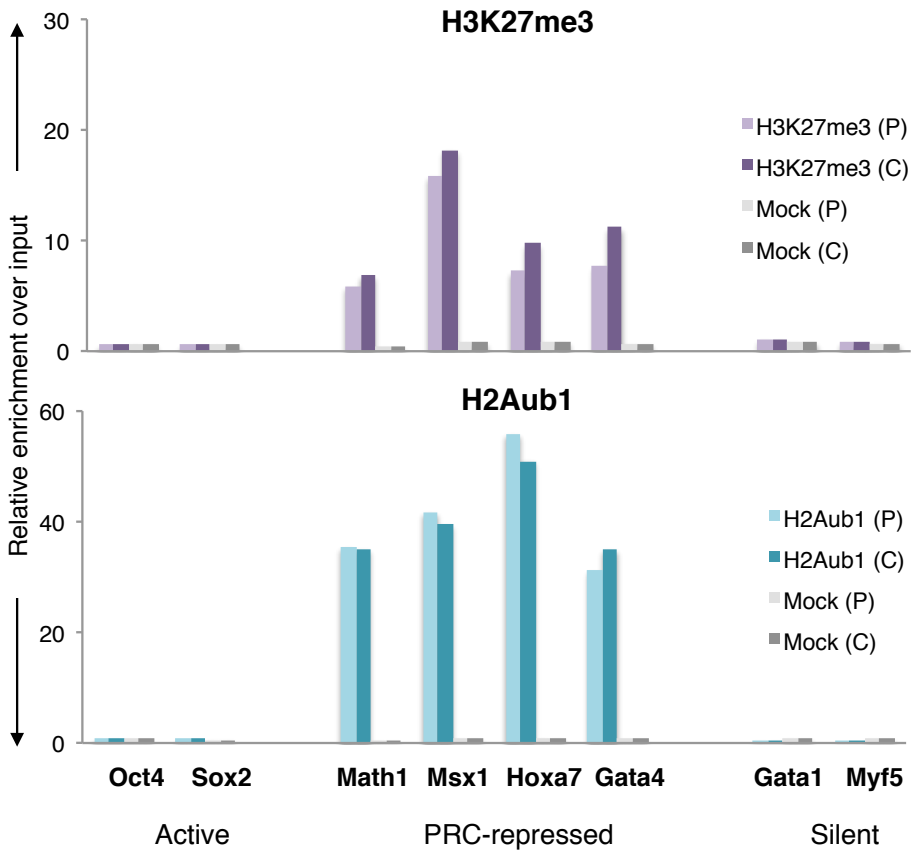


Figure 8.2. Occupancy of repressive histone modifications (H3K27me3 and H2Aub1) across a panel of active, PRC-repressed and silent genes on native chromatin in mES cells. Occupancy of H3K27me3 (light and dark purple) and H2Aub (light and dark cyan) as measured by DNA-ChIP on native chromatin and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Coding region primers map to ~2kb regions downstream of the TSS, except for Sox2 (-670bp). Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation with the Digoxigenin antibody (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. A single replicate is represented but additional samples were analysed with similar results.

8.4.4. H3K4me3 and H3K36me3

Next, I performed DNA-ChIP analysis of active histone modifications (H3K4me3 and H3K36me3) across a panel of genes in mES cells (Stock *et al.* 2007b; Brookes *et al.* 2012). H3K4me3 is a hallmark of open, accessible chromatin and marks actively promoters and PRC-repressed gene promoters (Jenuwein and Allis 2001; Brookes and Pombo 2009b). H3K36me3 is mark of transcriptional elongation catalysed by Set proteins and found at actively

transcribing gene bodies (Brookes and Pombo 2009b; Sims and Reinberg 2009). I performed DNA-ChIP using highly specific antibodies directed against H3K4me3 and H3K36me3 and observed H3K4me3 enrichment at promoters of active and PRC-repressed genes. H3K4me3 was comparatively lower at coding regions (+2kb downstream of TSS) than at promoter (Fig. 8.3). H3K36me3 was found enriched at coding regions of only active genes and not at PRC-repressed genes consistent with role as mark of transcriptional elongation (Fig. 8.3). Mock ChIP demonstrates the specificity of our ChIP protocol, with no detectable enrichment across active, PRC-repressed or silent genes.

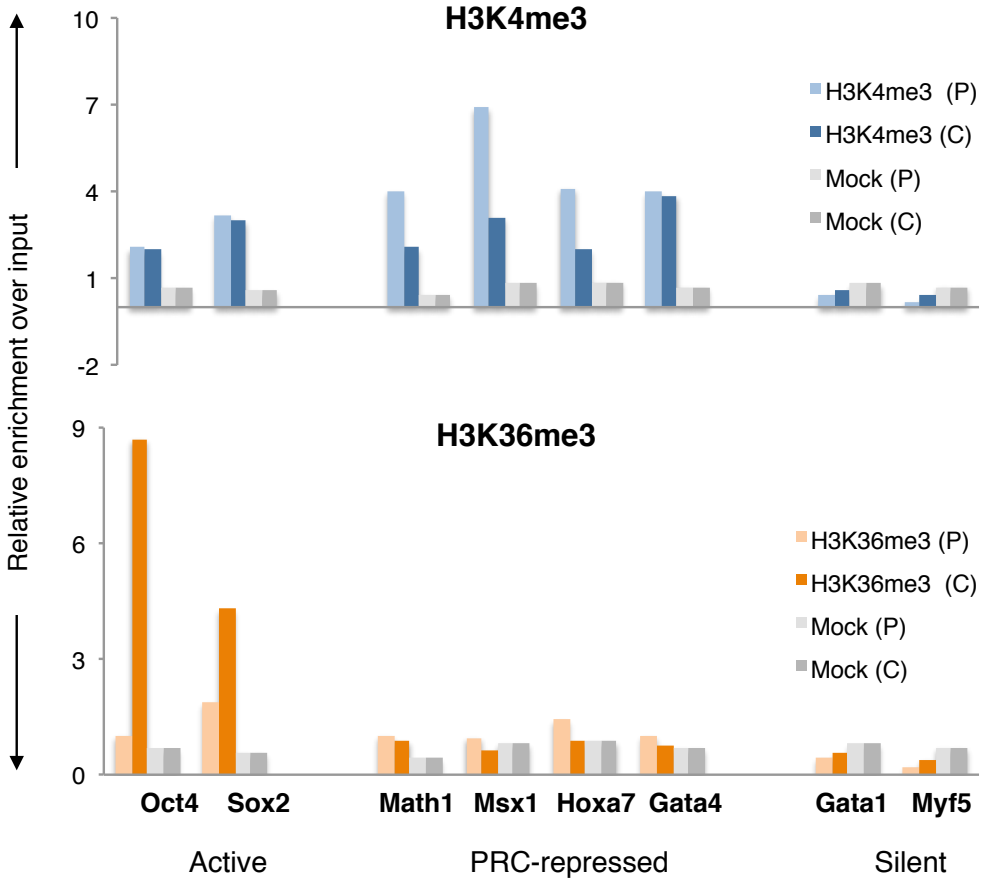


Figure 8.3. Occupancy of active histone modifications (H3K4me3 and H3K36me3) across a panel of active, PRC-repressed and silent genes on native chromatin in mES cells. Occupancy of H3K4me3 (light and dark blue) and H3K36me3 (light and dark orange) as measured by DNA-ChIP on native chromatin and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Coding region primers map to ~2kb regions

downstream of the TSS, except for Sox2 (-670bp). Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation with the Digoxigenin antibody (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. A single replicate is represented but additional samples were analysed with similar results.

8.4.5. Ezh2 and Ring1b

Polycomb proteins Ezh2 and Ring1b catalyse repressive histone modifications H3K27me3 and H2Aub respectively. I next performed DNA-ChIP analysis for these two proteins across a panel of genes in mES cells (Stock *et al.* 2007b). In our results, Ezh2 is enriched promoters and coding regions (+2kb downstream of TSS) of only PRC-repressed genes (Fig. 8.4); consistent with H3K27me3 occupancy (Fig. 8.2). Ring1b follows a similar pattern and is enriched only at promoter and coding regions PRC-repressed genes (Fig. 8.4). Both Polycomb proteins were not enriched at active or silent genes. Mock ChIP demonstrates the specificity of our ChIP protocol, with no detectable enrichment across active, PRC-repressed or silent genes.

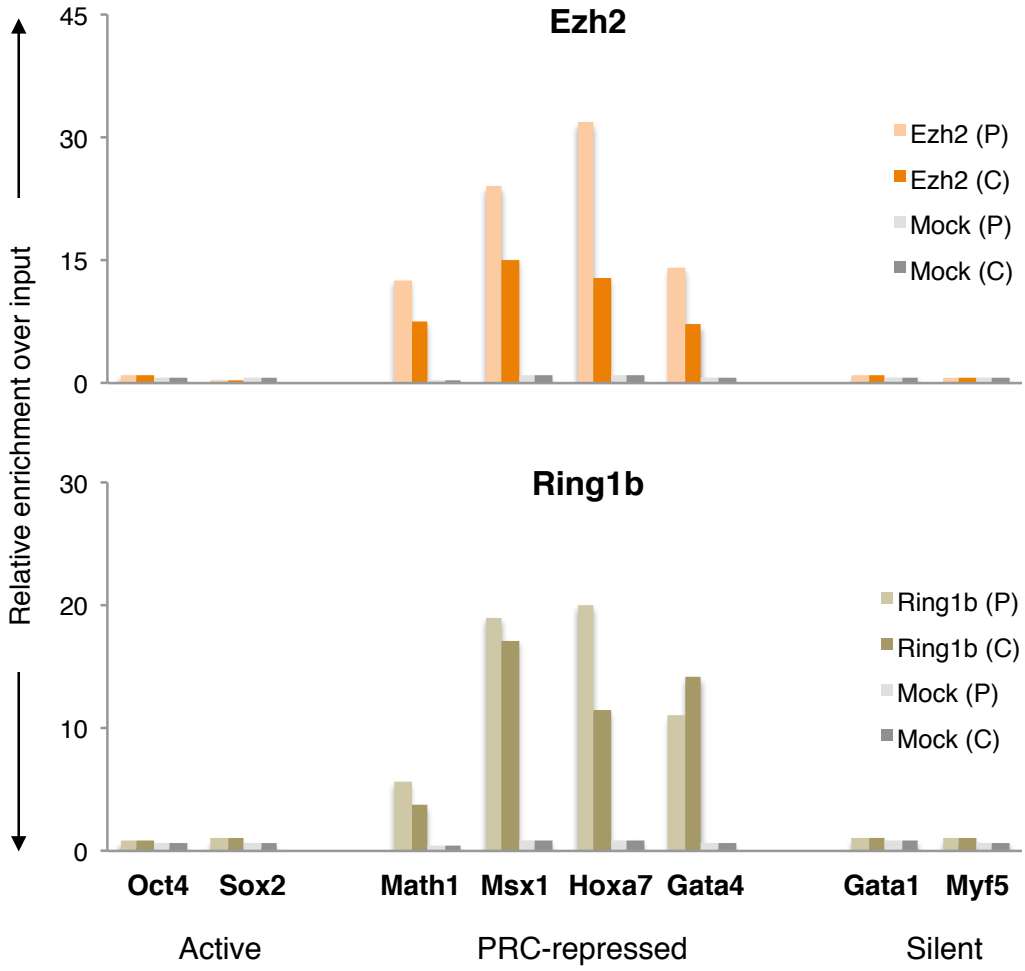


Figure 8.4. Occupancy of repressive Polycomb proteins (Ezh2 and Ring1b) across a panel of active, PRC-repressed and silent genes on native chromatin in mES cells. Occupancy of Ezh2 (PRC2 subunit; light and dark orange) and Ring1b (PRC1 subunit; light and dark brown) as measured by DNA-ChIP on native chromatin and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Coding region primers map to ~2kb regions downstream of the TSS, except for Sox2 (-670bp). Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation with the Digoxigenin antibody (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. A single replicate is represented but additional samples were analysed with similar results.

8.4.6. RNAPII modifications (S5p, S7p and S2p)

Native chromatin has much higher resolution spanning only mono- and di-nucleosomes and therefore DNA-ChIP with RNAPII is often not attempted, under the assumption that RNAPII is lost. In addition, it is known that RNAPII binds at the nucleosome free region (NFR) at +1 nucleosome and this

potentially limits the detection and capture of RNAPII using native chromatin. However, native chromatin followed by low salt extraction has been recently used to map RNAPII and nucleosome turnover in *D. melanogaster* (Teves and Henikoff 2011).

I next performed DNA-ChIP on native chromatin using our highly specific and optimised protocol (Stock *et al.* 2007b) to measure the occupancy of RNAPII modifications (S5p, S7p and S2p) in native chromatin. Consistent with crosslinked DNA-ChIP results (Figs. 3.6 and 4.4; (Stock *et al.* 2007b), RNAPII-S5p is enriched at promoters and coding regions of active and PRC-repressed genes in native chromatin (Fig. 8.5). RNAPII-S7p is enriched only at promoters and coding regions of active genes. RNAPII-S2p is also predominantly enriched at coding regions of active genes. Inactive genes had no enrichment for either RNAPII modification (Fig. 8.5). Mock pChIP was performed in parallel to all pChIP experiments and demonstrates minor levels of non-specific ChIP enrichment. We observe differential levels of RNAPII enrichment (S5p, S7p and S2p) between native ChIP and crosslinked ChIP due the different complexity and resolution of chromatin (Figs. 8.4 and 3.6). Our ability to detect RNAPII modifications using native chromatin highlights the sensitivity of our DNA-ChIP protocol and additionally allows capturing chromatin interactions at much lower resolution (than crosslinked chromatin) in mES cells. It is interesting to note that for S2p DNA-ChIP, DNA yields were proportionally lower in native than crosslinked ChIP.

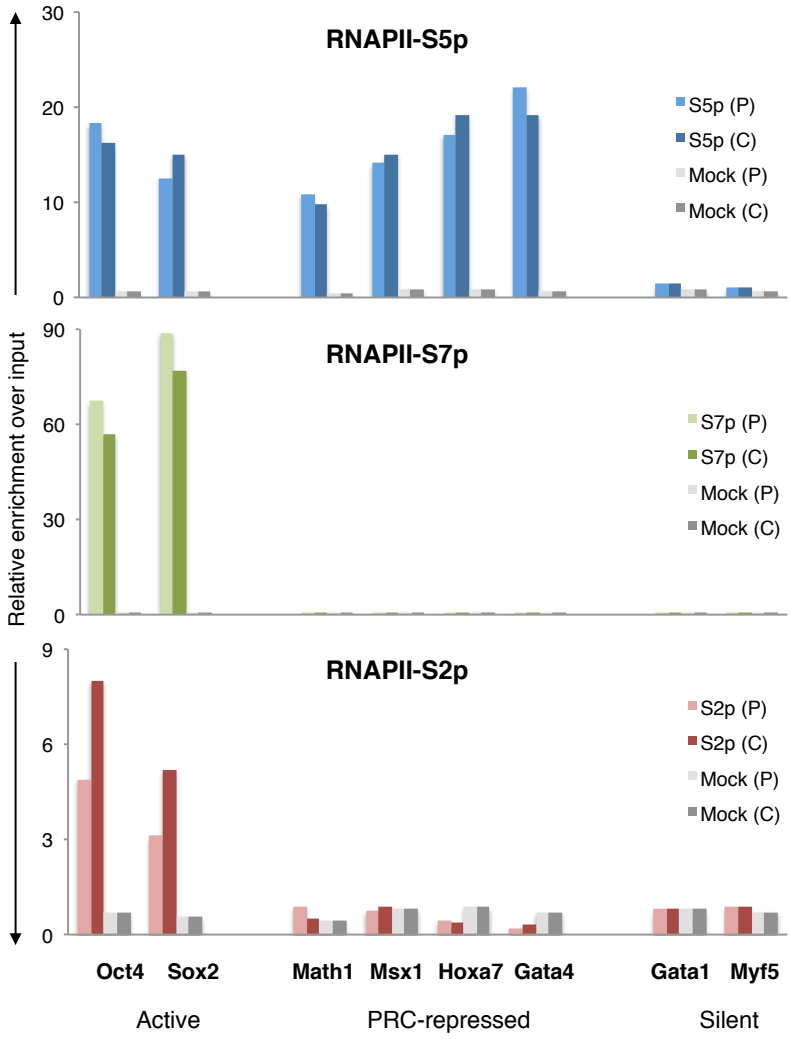


Figure 8.5. Occupancy of different RNAPII modifications across a panel of active, PRC-repressed and silent genes on native chromatin in mES cells. Occupancy of RNAPII-S5p (light and dark blue), -S7p (light and dark green) and -S2p (light and dark red) as measured by DNA-ChIP and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Coding region primers map to ~2kb regions downstream of the TSS, except for Sox2 (-670bp). Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation without primary antibody (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR. A single replicate is represented but additional samples were analysed with similar results.

8.4.7. Western blotting for RNAPII-S5p in pChIP samples performed on native chromatin.

After confirming the occupancy of RNAPII modifications (S5p, S7p and S2p) in native chromatin using DNA-ChIP, I next performed pChIP with RNAPII, histone modifications and Polycomb antibodies to check if the amount of protein extracted in each case could be visualised by western blotting. Sensitivity of western blot ranges from 1-10ng for most proteins and visualization aids in determining roughly protein amounts. Approximation of protein amounts also aids in selecting appropriate MS pre-processing steps and MS run-time albeit not essential for MS analysis.

After performing pChIP with RNAPII (S5p, S7p and S2p), histone modifications (H3K27me3 and H2Aub), Polycomb proteins (Ezh2 and Ring1b) and mock, I denatured and eluted the immunoprecipitated proteins by boiling the beads in custom Laemmli buffer (95°C for 10min) and removed the supernatant from beads. Protein samples were separated on 15% SDS-PAGE, followed by western blotting using anti-RNAPII-S5p antibody; Coomassie staining was not attempted to measure yield of protein extraction due to the low protein yields in pChIP and low sensitive of Coomassie. We observe that S5p is robustly enriched in multiple pChIP samples including S5p and S7p, consistent with the highest abundance of both marks on chromatin in these samples. S5p was also detected above control IP in the Ezh2 and Ring1b pChIP lanes after long exposure (Fig. 8.6; Lane 4,5,9 and 10). Low level of signal was observed in H3K27me3, H2Aub and S2p pChIP upon overnight exposure (Fig. 8.6; Lanes 6, 7 and 8; data now shown). Mock pChIP was clean and even on long exposure (overnight) no S5p signal is observed demonstrating the specificity of our pChIP and western blotting.
Elongating RNAPII is marked by both S5p and S2p as demonstrated by sequential ChIP analyses (Brookes *et al.* 2012). It is possible that in these brief preliminary analysis S2p pChIP has lower yields of proteins that prevent robust detection of S5p by western blotting. It is also interesting to notice the

lower S2p enrichments in native DNA-ChIP (Fig. 8.5). Future western blot analyses with S2p antibodies will be help understand the low RNAPII-S5p detection in the S2p pChIP sample.

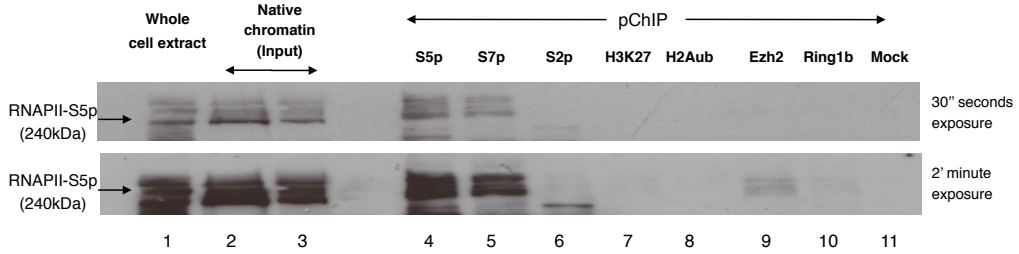


Figure 8.6. Western blotting confirms immunoprecipitation of RNAPII-S5p with different pChIP samples on native chromatin. Western blotting was performed using anti-RNAPII-S5p antibody on pChIP samples (S5p, S7p, S2p, H3K27me3, H2Aub, Ezh2, Ring1b and mock) and clearly show enrichment for S5p in S5p, S7p, Ezh2 and Ring1b pChIP samples (lanes 4,5,9,10). Lower levels of signal were obtained in other samples (S2p, K27me3, and H2AUb; Lanes 6,7 and 8) on long exposure. Mock (lane 11) was consistently devoid of S5p band. Proteins were denatured by boiling beads after pChIP at 95°C for 10min in custom Laemmli buffer and eluting supernatant from beads.

8.4.8. pChIP-MS on native chromatin

After confirming the occupancy of histone modifications, Polycomb proteins and RNAPII modifications on native chromatin, I next performed pChIP using a subset of antibodies (S5p, Ring1b, H3K27me3, H2Aub, H3K36me3) along with mock (Digoxigenin and no antibody) and analysed the proteins by MS. MS processing steps were performed by Dr. Bram Snijders (at Proteomics facility at MRC-CSC). We first removed common MS contaminants (keratins, immunoglobulins, etc.) and used both of our mock pChIP experiments to filter contaminants. As in this case, we have not used SILAC labelling we used standard MS scoring (Mascot score) for peptide and protein identification. Mascot score is probability based scoring that matches observed peptide mass values (from tandem MS/MS fragment ion masses) against a random event to identify significant, robust peptides. Any proteins with Mascot score > 30 in either mock pChIP (Digoxigenin or no antibody) were considered non-specific ChIP contaminants, and removed from the dataset. In addition, we

used further stringent criteria for each pChIP, *i.e.* mascot score > 30 to demarcate specific associated proteins from non-specific proteins. The pChIP were done on native chromatin and MS analysis was accordingly performed. Due to the non-quantitative nature of the analysis, we first looked at the total dataset to obtain overview of the MS run and then later performed each pChIP data analysis separately.

We identified 4655 proteins from the 6 different MS samples and included major groups of proteins including transcription factors, metabolic proteins, Zinc finger, Ring finger proteins and splicing factors. Table 8.1 shows the summary and number of proteins in each group. We also looked at different classes of enzymes important for regulation of several biological processes and saw a distinct enrichment for kinases, phosphatases and transferases. Table 8.1 lists the different classes of enzymes identified. Interestingly we observed about ~350 proteins enriched across all the different pChIP (except mock and Digoxigenin) with high mascot scores (>100) and involved in diverse processes. To keep our analysis unbiased, we did not filter these proteins and continued further analysis.

Table 8.1. Summary of proteins identified in pChIP on native chromatin. Five pChIP samples (S5p, Ring1b, H3K27me3, H2Aub and H3K36me3) and native input chromatin were analysed by MS. 4655 unique proteins were identified from the total dataset. Example of major classes of proteins and enzyme are listed below.

Total number of proteins in all pChIP dataset – 4655 proteins
(S5p, Ring1b, H3K27me3, H2Aub, H3K36me3 and native input chromatin)

Protein group	**Number of proteins**
Transcription factors	104
Metabolic proteins	212
Zinc finger proteins	184
Ring finger proteins	36
Splicing factors	45
Enzymes classes	
All kinases	363
Serine/Threonine kinases	102
Phosphatases	99
Methyl transferases	57
Acetyl transferase	22

8.4.9. H3K27me3

H3K27me3 is a hallmark of repressive chromatin and is catalysed by Polycomb protein Ezh2. After performing H3K27me3 native pChIP, we identified 816 proteins. Comparing the depth of MS run, I first asked how many proteins were specifically enriched over input native chromatin (Fig. 8.7). H3K27me3 pChIP specifically enriched for 278 proteins (overlap – 34% of H3K27me3 proteome). Gene ontology (GO) analysis identifies terms ‘nucleosome disassembly’, ‘chromatin modification’, ‘ncRNA processing’ and ‘cell cycle’ (Fig. 8.7). We identify additional proteins in H3K27me3 only (538 proteins) that reflect on incomplete protein sequencing depth of input

chromatin. A few examples of proteins enriched in H3K27me3 are listed in the Fig. 8.7.

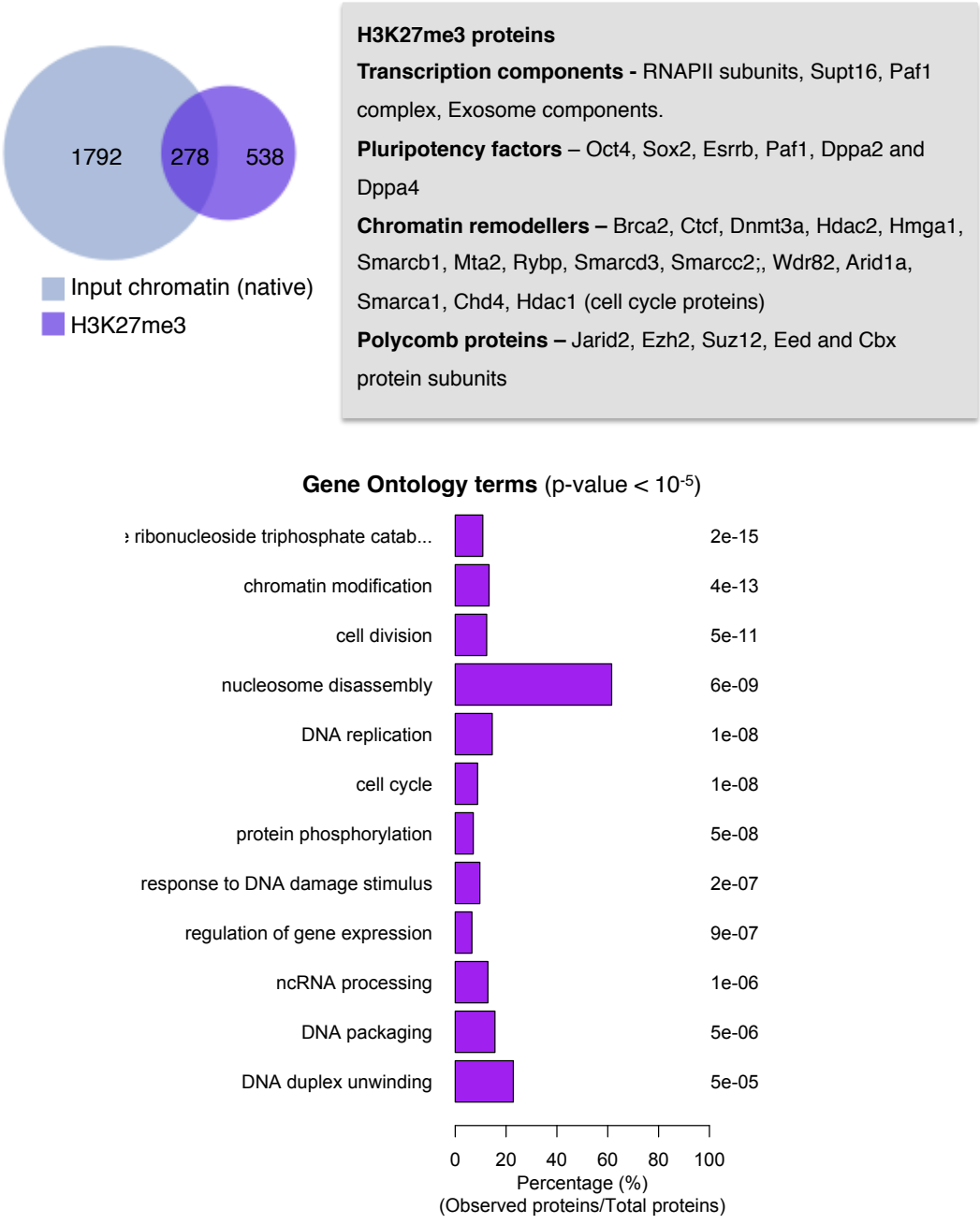
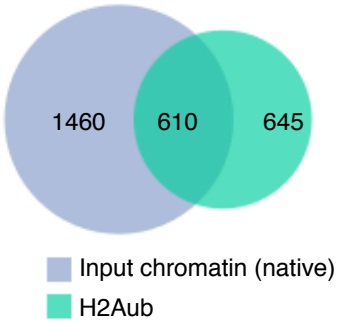


Figure 8.7. Summary of proteins identified in H3K27me3 pChIP using native chromatin. Proteins identified in H3K27me3 native pChIP were contrasted with input native chromatin proteins as plotted as Venn diagram. Examples of proteins enriched in H3K27me3 are also represented and include transcriptional components, chromatin remodellers and Polycomb proteins. Significant GO terms enriched and their p-values are represented as purple bar plots.

8.4.10. H2Aub

H2Aub is also a repressive mark, catalysed by protein Ring1b (PRC1 subunit). Performing H2Aub native pChIP, we identify 1255 proteins and comparing the depth over input native chromatin, H2Aub pChIP specifically enriches for 610 proteins (overlap – 49% of H2Aub proteome). Performing gene ontology (GO) analysis, terms enriched with H2Aub include ‘DNA metabolic processes’, ‘chromatin modification’, ‘gene silencing by miRNA’ and ‘negative regulation of expression’ (Fig. 8.8). Similar to H3K27me3 pChIP, additional proteins are identified in H2Aub only (645 proteins) reflecting low sequencing depth of input chromatin.



H2Aub proteins

Polycomb proteins – Jarid2, Ezh2, Suz12, Rbbp4/7 and Eed

Transcription components - RNAPII subunits, Paf1 complex ,Mediator subunits, splicing factors, Exosome subunits, Drosha, Tarbp2

Chromatin remodellers – Chd1, Hdac2, Dnmt3a, Mbd3, Hdac7, Wdr5, Suvar420h2, Aebp2, Phf8, Phf17 and Kdm3a

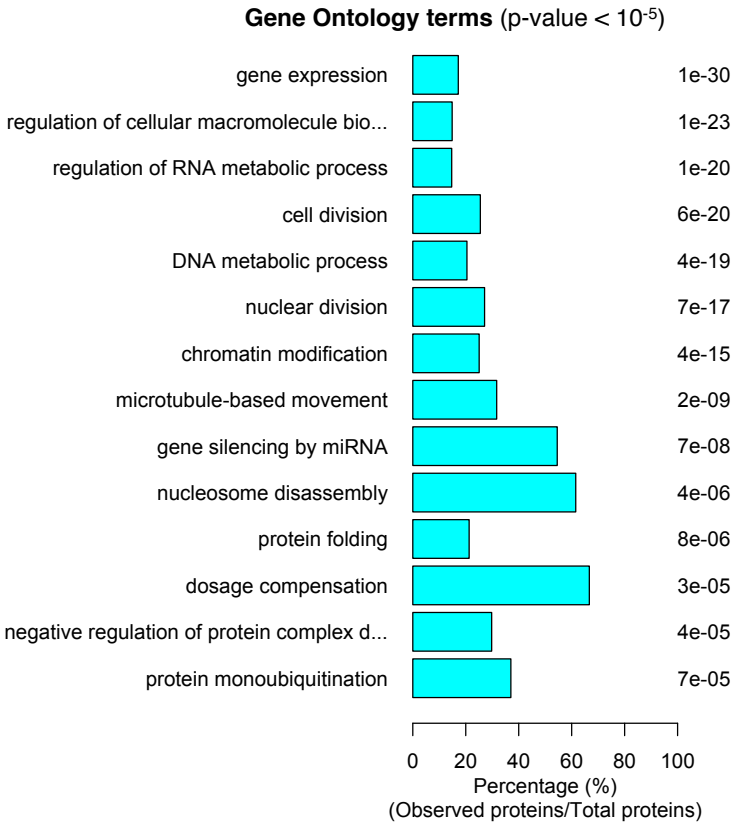
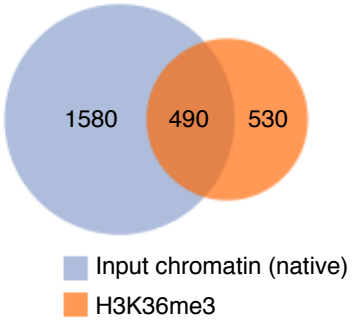


Figure 8.8. Summary of proteins identified in H2Aub pChIP using native chromatin. Specificity and robustness of H2Aub native pChIP was analysed by contrasting with input native chromatin proteins and plotted as Venn diagram. Examples of proteins enriched in H2Aub are also represented and include Polycomb proteins, some transcriptional components and chromatin remodellers. Significant GO terms enriched and their p-values are represented as cyan coloured bar plots.

8.4.11. H3K36me3

H3K36me3 is present in gene bodies and is hallmark of transcriptional elongation characterised by open chromatin. From the H3K36me3 native pChIP, we identified 1020 proteins. Comparing the overlap with proteins from input native chromatin, we identify 490 proteins (overlap – 48% of H3K36me3) enriched over input chromatin, Performing gene ontology (GO) analysis, enriched terms included ‘RNA processing’, ‘GTP metabolic processes’, ‘mRNA transport’, ‘chromatin modification’ terms and ‘mRNA catabolic processes’ (Fig. 8.9). Consistent with low input native chromatin sequencing depth, we identified 530 proteins in H3K36me3 only (912 proteins). Examples of proteins enriched in H3K36me3 are listed in the Fig. 8.9.



H3K36me3 proteins

Transcription components - RNAPII subunits, Splicing machinery (Cstf and Cpsf subunits), Ddx and Dhx proteins, Rest compelx subunits, Exosome subunits, Tho-Trex subunits, Ncbp1

Chromatin remodellers –Baz1b, Mbd2, Cdk9, Kdm5b, Rybp, Mta2, Oct4

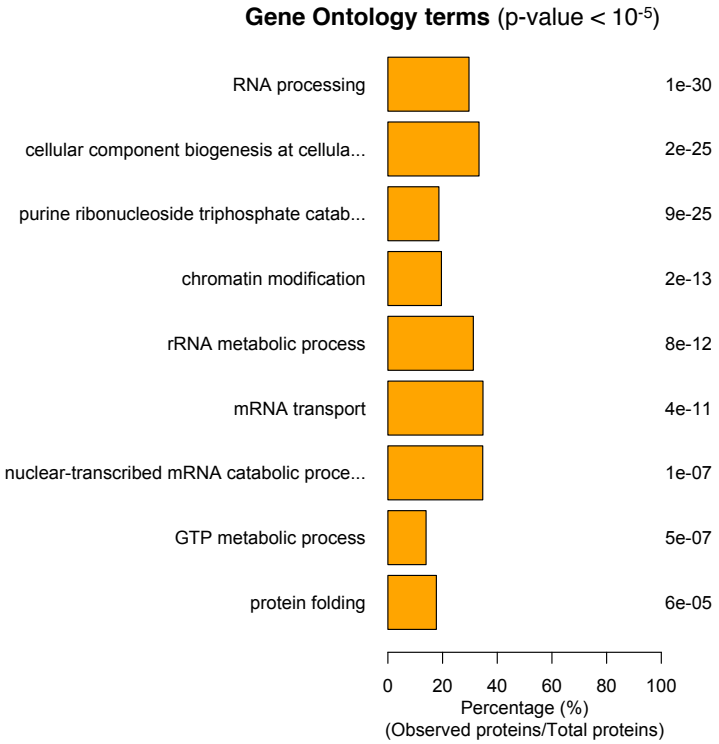
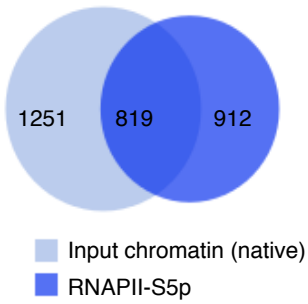


Figure 8.9. Summary of proteins identified in H3K36me3 pChIP using native chromatin. Proteins identified in H3K36me3 native pChIP were first contrasted with input native chromatin proteins and plotted as Venn diagram. Examples of enriched are also represented. Significant GO terms enriched and their p-values are represented as orange coloured bar plots.

8.4.12. RNAPII-S5p

From the RNAPII-S5p native pChIP, we identified 1731 proteins which include ~350 proteins having high mascot scores across all pChIP experiments. I first asked whether RNAPII-S5p pChIP specifically enriches for proteins from input

native chromatin. Using the MS with intermediate depth, RNAPII-S5p pChIP specifically enriched for 819 proteins of 2070 proteins detected in input chromatin (overlap – 47% of RNAPII-S5p proteome). Performing gene ontology (GO) analysis, I observed that the proteome enriched for with RNAPII-S5p pChIP was enriched for 'mRNA metabolic processes', 'transcription', 'gene silencing ', 'chromatin modification' terms and 'blastocyst formation' (Fig. 8.10). The identification of additional proteins in RNAPII-S5p only (912 proteins) reflects on incomplete protein sequencing depth of the more complex input chromatin. A few examples of proteins enriched in RNAPII-S5p are listed in the Fig. 8.10.



S5p only proteins

Transcription components - RNAPII subunits, Splicing machinery (Cstf and Cpsf subunits), Ddx and Dhx proteins.

Pluripotency factors – Oct4, Sox2, Esrrb, Paf1, Dppa2 and Dppa4

Chromatin remodellers – Aebp2, Hdac2, Hdac6, Prmt1, Mecp2, Kdm4a, Kdm6, Baz2a, Arid1a, Smarcad1/2, Setdb2

Polycomb proteins – Jarid2, Ezh2, Suz12, Eed and some cbx protein subunits

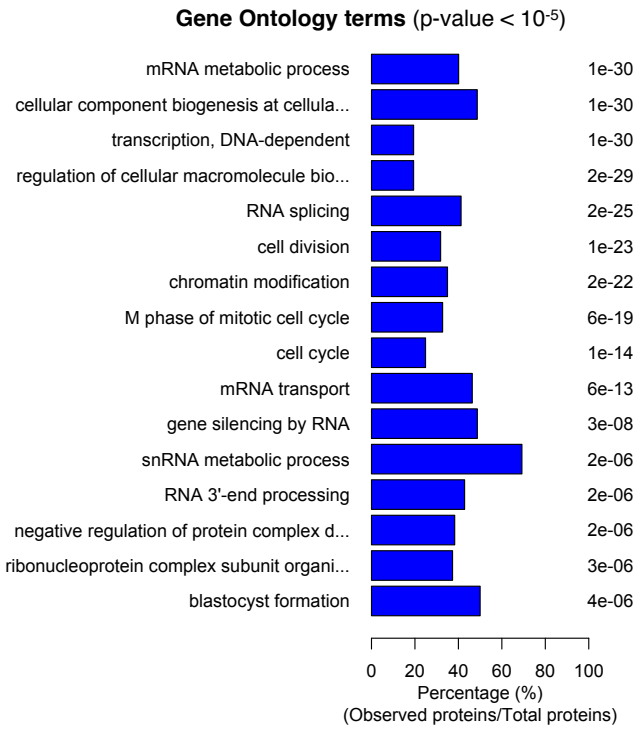
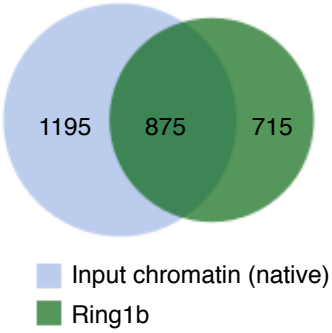


Figure 8.10. Summary of proteins identified in RNAPII-S5p pChIP using native chromatin. Proteins identified in RNAPII-S5p native pChIP were contrasted with input native chromatin proteins as plotted as Venn diagram. Examples of proteins enriched in RNAPII-S5p are also represented. Significant GO terms enriched and their p-values are represented as barplots.

8.4.13. Ring1b

Ring1b is a Polycomb protein (PRC1 subunit) and performing Ring1b native pChIP, we identified 1590 proteins. Comparing the depth of MS run, I first asked how many proteins were specifically enriched over input native chromatin (Fig. 8.11). Ring1b pChIP specifically enriched for 875 proteins

(overlap – 56% of Ring1b proteome). Performing gene ontology (GO) analysis, terms enriched with Ring1b included 'Ribonucleoprotein complex biogenesis', 'ATP metabolic processes', 'chromatin modification' and 'gene silencing' (Fig. 8.11). Similar to RNAPII-S5p native pChIP, we identified additional proteins in Ring1b only (715 proteins) reflects on protein sequencing depth of input chromatin. A few examples of proteins enriched in Ring1b are listed in the Fig. 8.11.



Ring1b proteins

Polycomb proteins – Jarid2, Ezh2, Suz12 and Eed

Transcription components - RNAPII subunits, Paf1 complex,, Ddx and Dhx proteins, Exosome components (Exosc 2,3,4 5,9,10), Brd4

Pluripotency factors – Oct4, Esrrb, Paf1

Chromatin remodellers – Aebp2, Phf8, Phf17, Chd1, Brd4, Chd4, Hdac2, Hdac6, Arid1a

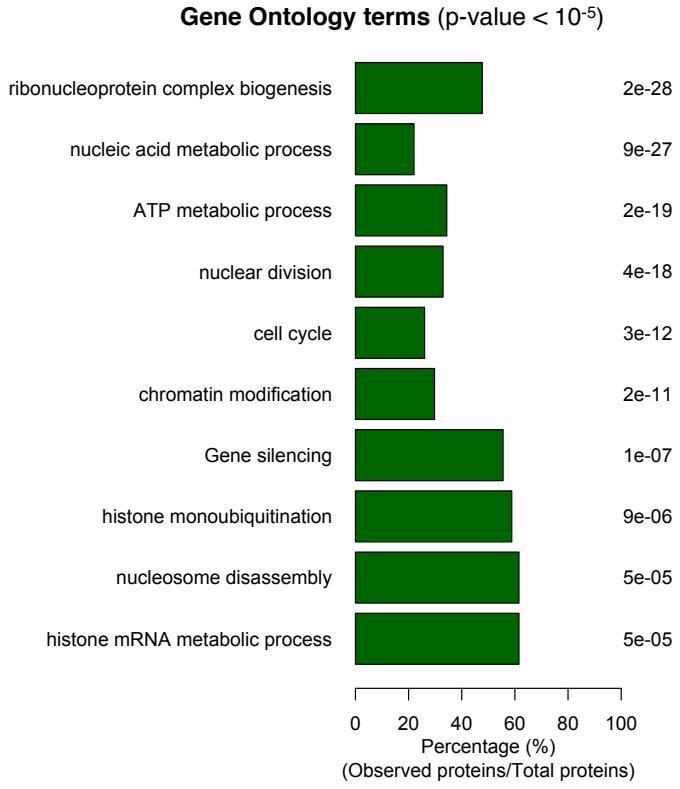


Figure 8.11. Summary of proteins identified in H2Aub pChIP using native chromatin. Specificity and robustness of Ring1b native pChIP was analysed by contrasting with input native chromatin proteins and plotted as Venn diagram. Protein examples include Polycomb proteins, quite a few transcriptional components, pluripotency factors and chromatin remodellers. Significant GO terms enriched and their p-values are represented as green coloured bar plots.

8.4.14. Gradient pChIP for crosslinked chromatin

Formaldehyde crosslinks all interactions within 2Å spacer arm including DNA-protein, protein-protein, RNA-protein and interactions dependent on chromatin states. With formaldehyde crosslinking, it is often taken into account that protein-protein interactions and further sub-interactions distant from chromatin are also crosslinked. In DNA-ChIP, these sub-interactions are less visible due to population effects and PCR amplification, whereas with pChIP, these interactions represent the pathway and range of protein-proteins interactions associated with the chromatin state chromatin state immunoprecipitated. Reducing these sub-interactions would facilitate the identification of most robust and direct chromatin interactors by pChIP in crosslinked chromatin. Therefore it is essential to purify the nucleo-histone complexes (only DNA, histones and interactors) without diluting/reducing the protein amounts while maintaining the sensitivity and detection limit.

8.4.15. Strategy for clarifying crosslinked chromatin to obtain nucleo-histone complexes and protein/DNA fractions

I first started by using the original DNA-ChIP protocol that included an additional step of using salt gradient to fractionate DNA-protein complexes from DNA-only and protein-protein only complexes (Gilmour and Lis 1984; Gilmour and Lis 1985; Solomon *et al.* 1988; Orlando *et al.* 1997). This additional step is removed from DNA-ChIP protocols primarily to save time without much effect on DNA yields (Schwartz *et al.* 2005). In addition, there have been reports that the analyses of specific gradient fractions can bias DNA-ChIP results for specific short DNA fragments (termed PRE-Polycomb Repressive Elements) in *D.melanogaster* (Schwartz *et al.* 2005).

Briefly, crosslinked chromatin was prepared as described (section 2.2.1.1). Density of fixed chromatin was adjusted to 567.8mg/ml (i.e. 1.42g/cm^3) using Cesium Chloride ($CsCl_2$) followed by ultra-centrifugation at high speed (40,000rpm, 72hrs, 4°C). The resulting gradient was collected in 10 (or 11) different fractions representing different densities of input chromatin, and excess salt from the fractions was subsequently removed by dialysis. After performing appropriate quality control experiments (see next section 8.3.7) and confirming good separation of nucleo-histone fractions from DNA-only and protein-only, DNA-ChIP and PChIP were accordingly performed using RNAPII antibodies (and other antibodies). After confirming the appropriate fractionation by different assays (see next section 8.1.7), DNA-ChIP (qRT-PCR) and pChIP (western blotting and MS) was performed using RNAPII antibodies to observe occupancy of RNAPII and its chromatin-bound interactome.

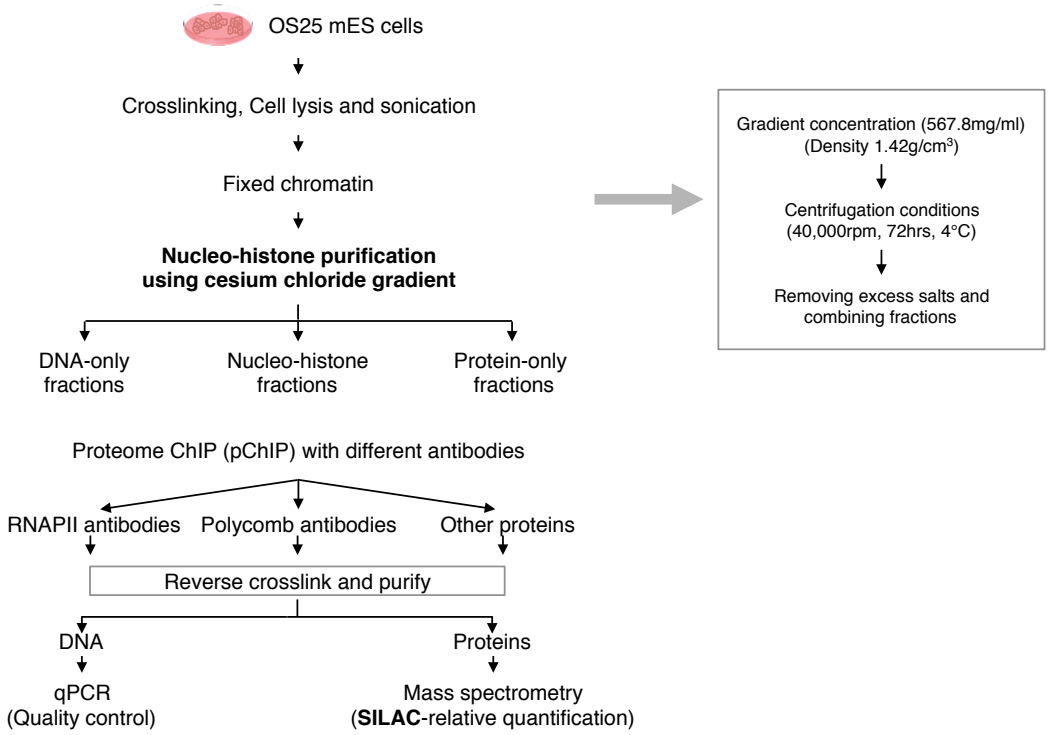


Figure 8.12. Overview of steps involved in Gradient-pChIP. Input chromatin was prepared as previously described (section 2.2.1.1). Input chromatin was fractionated

by salt gradient ($CsCl_2$) to obtain 10 different fractions representing DNA-only fractions, Nucleo-histone fractions and protein-only fractions (Schwartz *et al.* 2005).

8.4.16. Nucleo-histone complexes are well separated from DNA-only and protein-only fractions.

In the original DNA-ChIP protocols, the distribution of DNA (Agarose gel electrophoresis) and proteins (Coomassie staining or western blotting) across different fractions was collated along with density measurements (scintillation counter) to determine fractions containing nucleo-histone complexes while separating DNA-only and protein-only fractions (Gilmour and Lis 1984; Solomon *et al.* 1988; Orlando *et al.* 1997). The fractions were accordingly pooled together based on DNA and protein distributions for further ChIP assays and also for determining any biases either resulting from biological phenomenon (Schwartz *et al.* 2005) or inefficient gradient separation (Orlando *et al.* 1997).

8.4.17. Quality control for gradient fractions (Agarose gel, Coomassie and western blotting)

To optimise the gradient ChIP, I first compared the distribution of chromatin obtained from 11 different fractions on 1.2% Agarose gel (Fig. 8.13 Agarose gel). Loading a low volume of input chromatin (10µl), we observed a good gradient separation of chromatin across different fractions. Fractions 1-3 consisted of very short DNA fragments that migrated the fastest and were quite smaller. Fractions 4-8 consisted of bulk DNA, with maximum intensity in fraction 5. Fractions 9-11 consisted of very little DNA intensity and corresponding towards much bigger DNA fragments (>10kb; Data not shown).

Next, I observed the distribution of selected proteins by western blotting including 11 different fractions and unfractionated sample. Samples were first denatured by boiling samples in custom Laemmli buffer (95°C for 10min) and subsequently separated by electrophoresis (15% SDS-PAGE gel). Western blot using anti-S5p and anti-Oct4 antibody was performed (Fig. 8.13; Western

blot). RNAPII-S5p was robustly identified and enriched in the nucleo-histone fractions (top band; Fractions 4-8) and highest intensity was observed in fractions 6 and 7 (similar to unfractionated sample Lane 2). We also observed low intensity of S5p in fractions 9-11. Very little or no RNAPII-S5p was observed in fractions 1-3 (DNA-only fractions). Consistent with Agarose gel results (Fig. 8.13 A), we observed nucleo-histone complexes in fractions 4-8, DNA-only in fractions 1-3 and protein-only in fractions 9-11. We observed very faint blotting for Oct4 protein and Oct4 protein was predominantly enriched in nucleo-histone fractions (fractions 4-8).

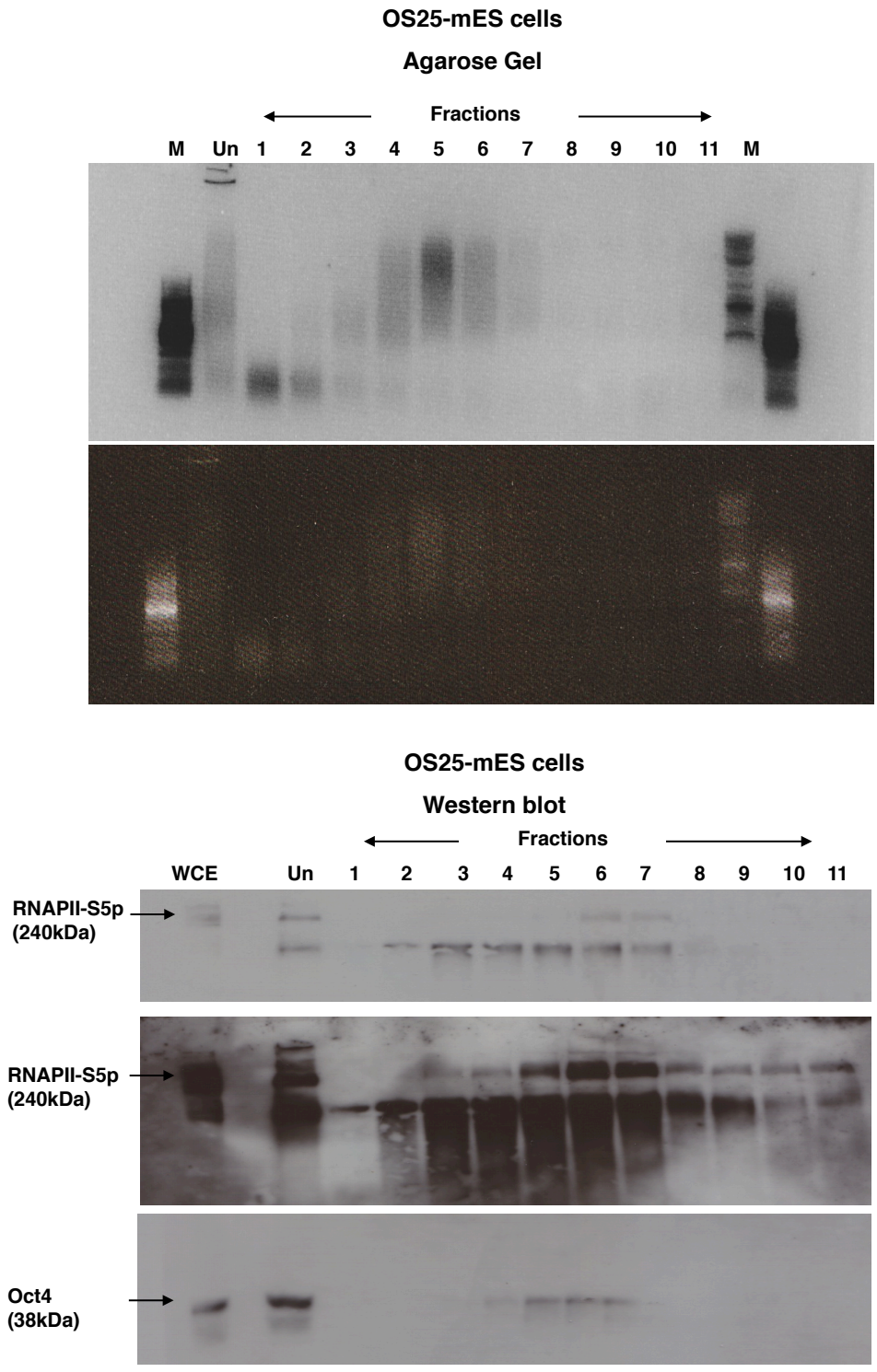


Figure 8.13. Distribution of DNA and western blotting after salt gradient fractionation. Eleven different fractions representing range of densities were obtained after $CsCl_2$ gradient separation. 10µl of different DNA fractions (after reverse-crosslinking) were loaded on 1.2% Agarose gel to observe the distribution of DNA. 10µl of reverse crosslinked protein sample (denatured at 60°C o/n and 100°C

10min in custom Laemmli buffer) was loaded on 15% SDS-PAGE gel to perform western blotting with anti-S5p antibody and anti-Oct4 antibody to observe protein distribution.

I next performed a basic visual comparison of the DNA and protein distribution from our OS25-mES cells to the distribution of *D.melanogaster* - Schneider L2 cells being aware of the differences in genome, transcriptome, proteome and their regulation between both organisms (Schwartz *et al.* 2005). Remarkably the distribution of gradient separation was similar between the two Agarose gels (Figure 8.13 and 8.14). The bulk DNA in both the cells was enriched across middle fractions (4-8 in our cells and densities-1.48-1.30 in Schneider cells).

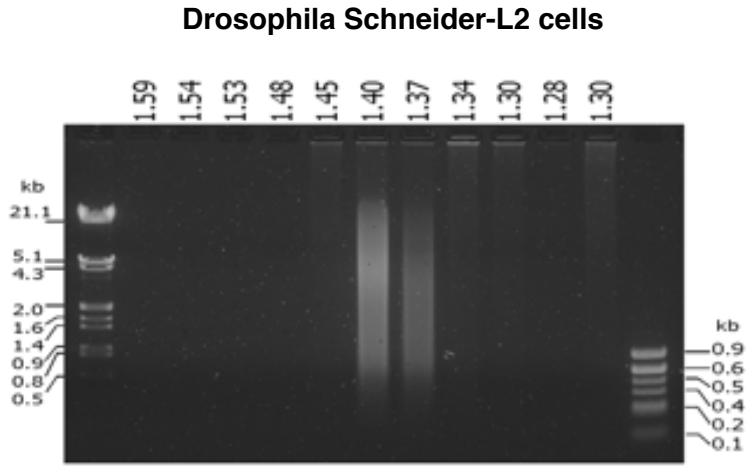


Figure 8.14. Similar distribution of DNA (chromatin) densities observed between *D.melanogaster* (Schwartz *et al.* 2005) and our mES cell chromatin. Brief comparison between published *D.melanogaster* DNA distribution across different density gradient fractions and our mES cell chromatin gradient fractions.

8.4.18. Distribution of active, PRC-repressed and silent genes across the different fractions.

I next looked at the distribution of specific genes across the different fractions to ensure that bulk nucleo-histone fractions contained most of the chromatin regions. Reverse crosslinking the input chromatin from different fractions and purifying the DNA, I next compared the distribution of two active, four PRC-repressed and two silent gene promoters across the 10 different fractions

using qRT-PCR (Fig. 8.15). In principle, all gene promoters are similarly enriched in unfractionated chromatin sample, therefore after fractionation gene promoters should only be enriched in nucleo-histone fractions. Consistently we observe fractions 4,5,6,7 and 8 containing bulk of the DNA (chromatin) and have similar pattern of enrichment for different gene promoters (active, PRC-repressed and silent; Fig. 8.15). X-axis represents the different fractions from 1-10; Y-axis represents relative enrichment (arbitrary units) and is calculated by using CT values for each fraction normalized to CT values for fraction 1.

It has been reported that gradient fractionation in *D.melanogaster* Schneider L2 cells biases DNA-ChIP results more specifically for short fragments called Polycomb Response Elements (PRE) (Schwartz *et al.* 2005). PREs have only been identified in *D.melanogaster* as short motifs that act as platform for Polycomb proteins to bind and regulate the chromatin architecture (Ringrose and Paro 2007; Pirrotta and Li 2012). It was reported that binding of Polycomb and accessory proteins on small PRE motif dynamically shifts the local density and upon gradient centrifugation PRE motifs are well detected across nucleo-histone fractions (Schwartz *et al.* 2005). In our mES cells, we didn't observe any bias for PRC-repressed genes (promoter region primers) or for active or silent genes (Fig. 8.15). I also looked for coding region primers for active, PRC-repressed, silent genes and observed no bias (data not shown).

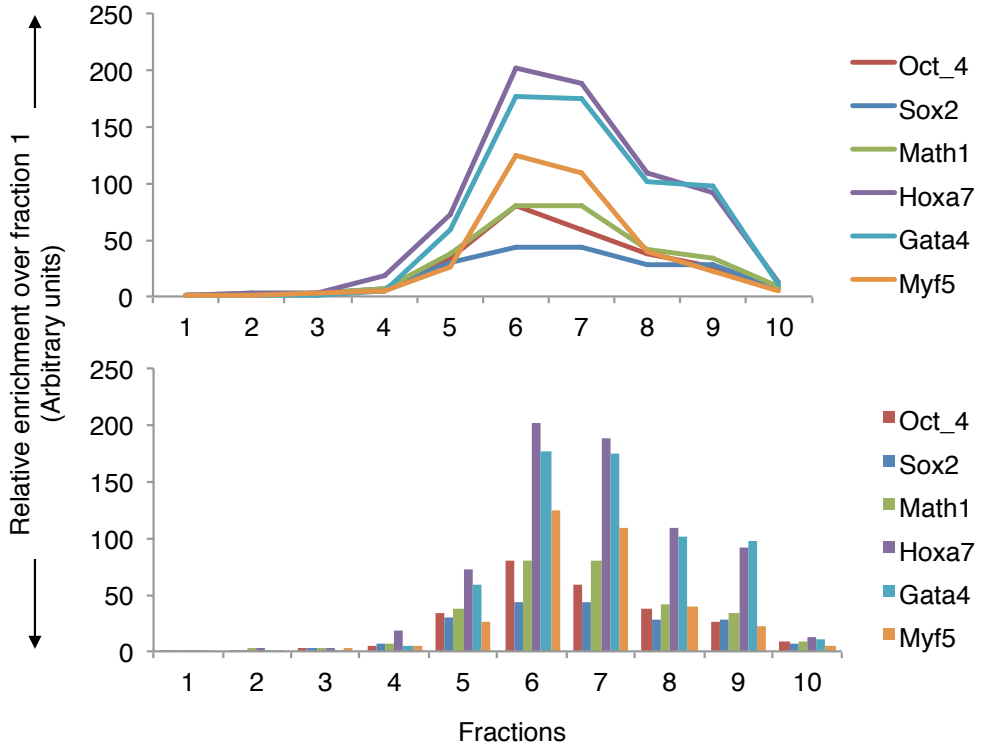


Figure 8.15. Confirming the abundance of different DNA regions preferentially in nucleo-histone fractions. DNA was obtained by reverse crosslinking chromatin and qRT-PCR was performed for 10 different chromatin fractions using primers spanning Promoter regions of two active, four PRC-repressed and two inactive genes. Nucleo-histone fractions (4,5,6,7 and 8) consisted of most of the DNA for all primers, very little DNA was observed in protein-only fractions 9 and 10. Pattern of enrichment for all primers was consistent across the different fractions. Differential levels of enrichments (Y-axis bars) are representative of different DNA amounts of the genes in the different fractions. Relative enrichment was calculated relative to fraction 1 and y-axis represents arbitrary units of enrichments.

8.4.19. Considerations for Gradient pChIP

To perform DNA-ChIP on gradient chromatin, we pooled fraction 1-3 together to form a DNA-only fraction, Fractions 4-8 to form a nucleo-histone fraction and fractions 9-11 to form a Protein-only fraction. To perform DNA-ChIP, we first quantified the concentration of DNA from 3 different pooled fractions (DNA yields ranging 10-50 µg). DNA concentration was used as a measure of chromatin volume for DNA-ChIP and pChIP.

8.4.20. DNA-ChIP for RNAPII modifications on gradient samples.

To observe the distribution and enrichment of RNAPII between the three different fractions, I first performed RNAPII-S5p DNA-ChIP on the three different fractions across a panel of promoter (P) and coding region (C) primers of two active, four PRC-repressed and two silent genes (Fig. 8.16). Robust enrichment of RNAPII-S5p is only observed in nucleo-histone fractions (Fig. 8.16; middle graph) and RNAPII-S5p is enriched at promoters and coding regions of active and PRC-repressed genes. These enrichments are consistent with previous results (Figs. 3.6 and 4.4) (Stock *et al.* 2007b; Brookes *et al.* 2012).

There was very little signal observed for RNAPII-S5p in DNA-only fractions with no significant enrichment (levels similar to mock DNA-ChIP). These results are consistent as input chromatin (of DNA-only fractions) had much lower starting material compared to nucleo-histone fractions (also see Fig. 8.15). We observed high levels of background (mock) in the Protein-only fractions and RNAPII-S5p enrichment levels were comparable to the background levels (Fig. 8.15 bottom panel; Grey bars). These results validate the specific enrichment of bulk chromatin in the nucleo-histone fractions and use of nucleo-histone fractions for further analysis.

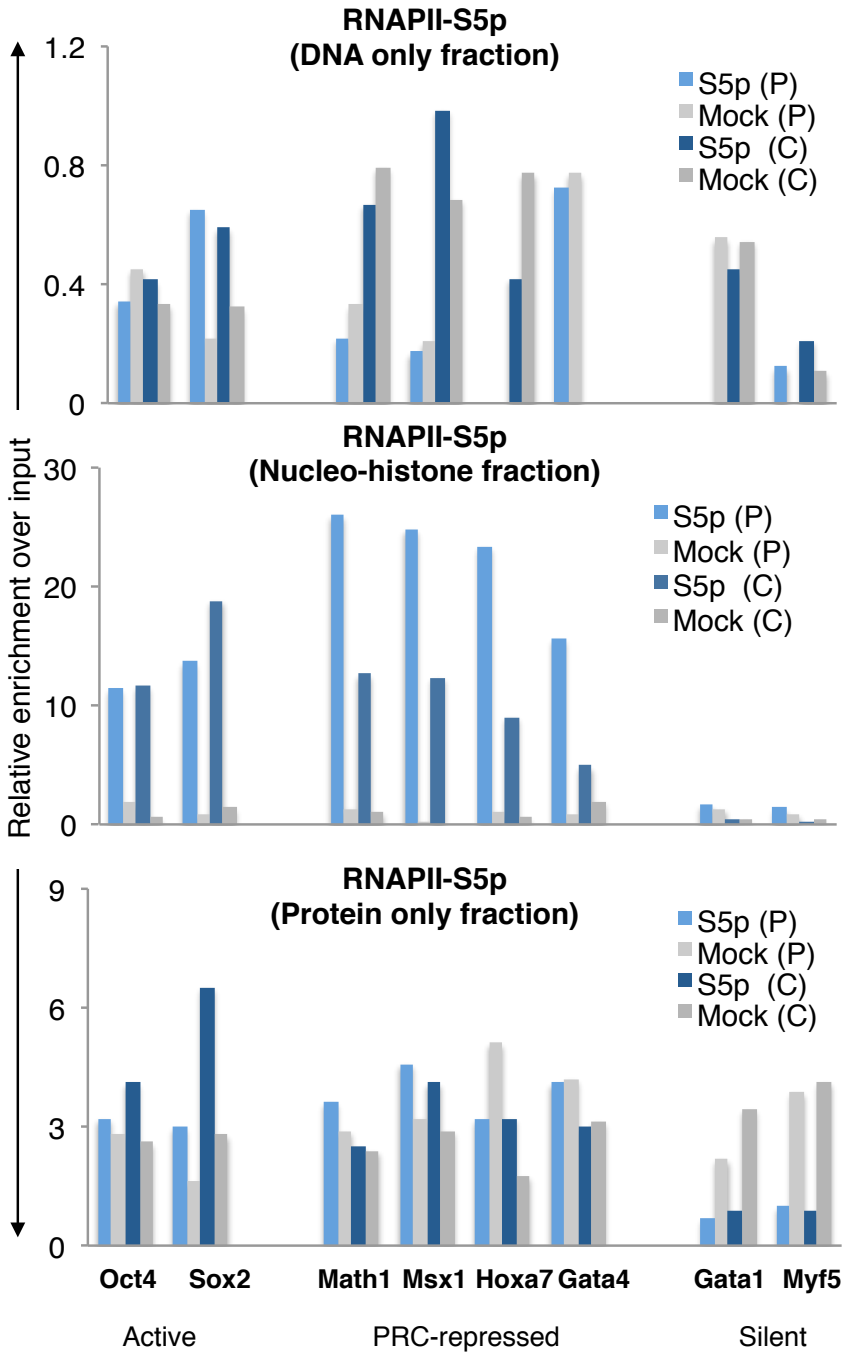


Figure 8.16. Occupancy of RNAPII-S5p DNA-ChIP measured on three chromatin fractions by gradient ChIP. Occupancy of RNAPII-S5p (light and dark blue was measured on three different chromatin fractions (DNA-only, nucleo-histone, Protein-only) by gradient DNA-ChIP and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR.

To compare the occupancy of other RNAPII modifications, I performed DNA-ChIP with RNAPII-S7p and S2p on the nucleo-histone fractions. Consistent with previous results (Figs. 3.6 and 4.4), the patterns of enrichment for RNAPII modifications are quite similar (Stock *et al.* 2007b; Brookes *et al.* 2012). RNAPII-S7p is enriched at promoter and coding regions of active genes and no enrichment is observed across PRC-repressed or silent genes. RNAPII-S2p is also enriched at active genes with higher levels of enrichment in coding regions compared to promoters.

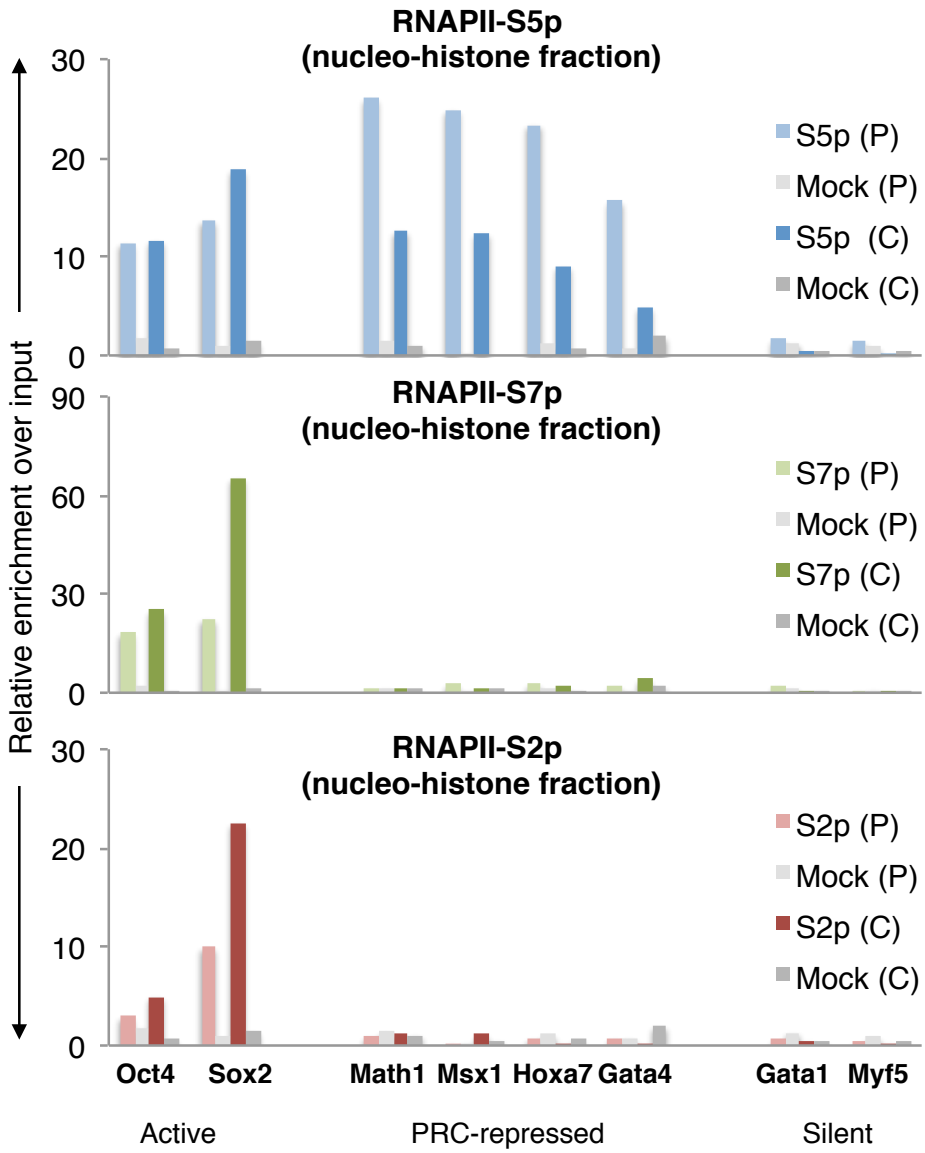


Figure 8.17. Occupancy of different RNAPII modifications across a panel of active, PRC-repressed and silent genes in nucleo-histone fraction by gradient DNA-ChIP in mES cells. Occupancy of RNAPII-S5p (light and dark blue), -S7p (light and dark green) and -S2p (light and dark red) as measured by DNA-ChIP and qRT-PCR at promoters (P) and coding (C) regions of panel of two active, four PRC-repressed and two inactive genes. Light and dark grey bars represent background enrichment levels as measured by control immunoprecipitation (Mock). Enrichment is expressed relative to input DNA using total amount of DNA in qRT-PCR.

8.4.21. MS analysis of pChIP gradient samples and next steps

I have performed pChIP with RNAPII modifications (S5p, S7p and S2p) on nucleo-histone fractions and these samples are currently being analysed by MS (Guido Mastrobuoni and Stefan Kempa at MDC-Berlin). In addition we are also processing RNAPII-S5p pChIP done on DNA-only and protein-only fractions to quantify the differences in protein identified.

In parallel, we are performing DNA-ChIP and pChIP experiments on nucleo-histone fractions using RNAPII modifications and Polycomb proteins to identify candidate proteins important for RNAPII-Polycomb interplay in mES cells.

8.5. Discussion

8.5.1. Diversity of native chromatin and pChIP proteins

PRC-repressed chromatin is marked by bivalent chromatin modifications along with RNAPII-S5p and Polycomb proteins that catalyse repressive histone marks. We have performed DNA-ChIP for histone modifications, Polycomb proteins and RNAPII modifications on native chromatin in mES cells. Reassuringly, the pattern of enrichment for histone modifications, Polycomb and RNAPII modifications are similar between crosslinked chromatin and native chromatin that suggests the specificity of our DNA-ChIP and chromatin preparation. We believe that there are some important considerations for robust DNA-ChIP in native chromatin in mES cells. Firstly, the extent of nucleosomal preparation *i.e.* micrococcal nuclease treatment is essential and requires optimisation. We observe that 5-min digestion (2U/ml MNase) robustly gives bands corresponding to primarily mono-nucleosomes and di-nucleosomes and rarely any poly-nucleosomes. Secondly, a well-characterised antibody against non-transient chromatin interactor is essential for a robust DNA-ChIP signal. Thirdly, we observe a good consistent yield of

DNA in native chromatin across different replicates that further add robustness to our protocol.

The diversity of native chromatin proteins has already been highlighted and discussed in Chapter 3. We observe an average of >2000 proteins detected from native input chromatin consistently across different runs in intermediate depth MS run. We observe a distinct enrichment for proteins with known association on chromatin including master stem cell regulators, TFs and several enzymes that modify histones and other chromatin proteins. Our pChIP samples analysed by MS were pooled from two immunoprecipitations done on same chromatin. The pooling of samples allows greater, better protein identification and enhances the MS peptide sequencing depth.

We have demonstrated the specificity of our pChIP and associating proteins on chromatin with different histone modifications, Polycomb proteins and RNAPII modifications. In our analysis, we observed ~300 proteins that were consistently detected across all pChIP runs (except mock). A naïve interpretation is that these proteins are non-specific due to their abundance in all experiments. However, the other possibility is that chromatin is marked by these proteins consistently (with subtle difference in stoichiometry or composition). Interestingly, important chromatin regulators belong to this group. The pChIP analysis on native chromatin was performed on normal chromatin and therefore was qualitative and not quantitative. Therefore, effective comparison between multiple pChIP runs was not possible. We strongly believe that in addition to quantitative analysis (e.g. SILAC) with complementary experimental setup, it is essential to perform systems biology analysis to unravel and detect patterns of protein associations to different proteins.

8.5.2. Gradient separation of chromatin fractions

In conclusion, I have adapted and optimised conditions for gradient fractions and purification of nucleo-histone fractions for our mES-OS25 cells. Performing quality control experiments, I observed enrichment of DNA in fractions 4-8 consistent with enrichment for histones and other proteins (Fig. 8.13 and data not shown) by Coomassie staining of proteins. From our preliminary analysis of chromatin, we did not observe any biases as reported with *D.melanogaster* cell lines (Schwartz *et al.* 2005). Since promoters of PRC-repressed genes have high occupancy of RNAPII-S5p, Polycomb proteins and bivalent histone modifications, we hypothesised that density gradient bias would be most apparent in differently sized PRC-repressed gene promoters. Testing promoter primers on input chromatin for two active, four PRC-repressed and two inactive genes, we observed no specific enrichment for either of the gene promoters. We cannot completely exclude the effect of fractionation without performing genome-wide analysis on different fractions.

We have also demonstrated the robustness of our nucleo-histone fractionation and confirmed specificity by performing DNA-ChIP on all fractions (DNA-only, protein only and nucleo-histone fractions). The gradient fractionation (DNA-ChIP and pChIP) offers specific advantages including increased yields of DNA and protein after immunoprecipitation. Secondly, non-specific background in both DNA and pChIP is massively reduced. Thirdly, due to fractionation, we enrich for chromatin state in our nucleo-histone fractions (compared to unfractionated sample) increasing the sensitivity and detection potential of pChIP-MS.

9. Discussion

9.1. Thesis overview

PRC-repressed genes encode for important developmental regulator, metabolic, signalling and lineage specification genes that exist with bivalent chromatin architecture along with Polycomb proteins and unusual RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) in mES cells. The chromatin state of these genes resolves into monovalent chromatin configurations upon lineage specification, suggesting the RNAPII and Polycomb proteins in mES cells maintain these genes in a poised state for future activation. In this thesis, I have investigated the chromatin landscape of mES cells from the point of view of RNAPII modification; I developed a novel unbiased strategy, called “Proteome-ChIP”, to explore the proteome associated with RNAPII-bound chromatin in mES cells.

The work in this thesis highlights the dynamic and complex nature of RNAPII gene regulation network in mES cells. In Chapter 3, I investigated the chromatin proteome extracted using different methods and their relevance to appropriate biological questions specifically in the context of stem cells. While optimising pChIP conditions, I have demonstrated that proteins associated with chromatin bound by RNAPII can be qualitatively and quantitatively enriched from input chromatin (Chapters 3 and 4). Using a combination of simple logics and advanced systems biology approach, we comprehensively dissect and unravel the RNAPII-chromatin bound proteome in mES cells and shed light on the dynamic RNAPII regulation and the complex integration of chromatin and RNA processing events that coincide on chromatin. From our datasets, we uncover novel RNAPII biology, S5p-associated processes and new knowledge of RNAPII specific stem cell processes (Chapters 5 and 6). I also investigated the versatility of pChIP to capture and uncover different interaction types (protein-protein, RNA-protein and DNA-protein) in Chapter 7. Lastly, extending and performing pChIP on native chromatin extracts and

fractionated chromatin, I have explored the robustness and potential of applying pChIP on different kinds of chromatin preparations to uncover novel chromatin association fundamental for control of gene regulation.

9.2. From research objectives to research findings

9.2.1. Proteome-ChIP as tool to unravel chromatin-bound proteome

Chromatin is composed of DNA wrapped round a histone octamer forming nucleosome monomers that are further associated by linker histone H1 establishing a higher level of chromatin organization (Fischle *et al.* 2003). The chromatin provides a platform for sequence-specific factors, chromatin remodellers, transcription factors, transcriptional machinery and other factors to access DNA and coordinate a range of regulatory and mechanistic programs that govern the cellular viability and response to external cues. This coordinated interplay on chromatin is tightly regulated and signals cascades of downstream processes that feedback and further regulate chromatin processes. To capture globally the proteome and many of these interactions, I initially used crosslinked chromatin extracts (Fig. 3.3) that chemically fixes and captures all types of interactions (DNA-protein, RNA-protein and protein-protein interactions); note that independent evidence shows that even newly-made RNAs are preserved and can be specifically purified after our RNAPII ChIP (K.J. Morris, our laboratory; unpublished work), implying that when interpreting the proteome isolated after ChIP, it is important to consider that some proteins may be immunoprecipitated through their association with RNA, and not necessarily through associations with chromatin or the RNAPII itself. To enrich specific interactions from all the possible interactions present in chromatin extracts, I performed pChIP using highly specific RNAPII antibodies (S5p and S7p: Fig 3.12 and Fig 3.13). Using high-throughput MS, we identify a large number of protein interactors by detecting unique and specific peptides corresponding to intact protein; this is an unique opportunity offered by MS detection, which unlike western blot analyses, does not require

detection or preservation of the whole protein, only that some peptides survive the assay until their ultra-sensitive detection by MS.

The ability to detect specifically and robustly chromatin-bound interactions is inherently linked to coupling the crosslinked chromatin preparation with my optimised pChIP protocol and aided by high-throughput detection of peptides by MS. In the literature, there have been other studies and methods to capture major protein interactions on chromatin including “PICh”, “mChIP” and “iChIP” (Dejardin and Kingston 2009; Hoshino and Fujii 2009; Lambert *et al.* 2009; Lambert *et al.* 2010). These methods have been quite useful in studying abundant proteins but have often required large amounts of starting material. In addition, issues concerning low sample recovery, sample heterogeneity, and impossibility of amplifying protein samples prior to their MS analysis affect these methods. Some of these methods require probes against specific DNA sequence or require random insertion of tagged constructs; that makes them less appealing for use in stem cells. We have demonstrated that coupling pChIP protocol to crosslinked chromatin and appropriate MS conditions, combined with enrichment for chromatin compartments associated with specific proteins (such as RNAPII) allows robust qualitative and quantitative detection of proteins, overcoming the issues of the other methods.

Another determinant in chromatin proteomics methods is low abundance of proteins reflected by weak Coomassie staining (Fig. 3.9) that often discourages further investigation into the proteins. However the sensitivity of MS to detect peptides (instead of whole proteins) and coupling with pChIP like methods allow identification and protein dependencies to RNAPII modifications. Tandem MS/MS provides qualitative information on proteins, *i.e.* whether a peptide is detected or not and the measure of its detection (Mascot score). In label-free proteomics, no inference can be reached for peptides that are not detected. Additionally label-free proteomics does not provide any information on quantitative aspects including relative abundance per cell or relative abundance to input and/or reference sample. This is a

major issue for low abundant proteins as between different runs of the same sample; peptides are often missed due to detection limits. Additionally, identified proteins (and their Mascot scores) cannot be cross-compared between different runs due to high variability and limited normalization measures. To tackle these issues, multiple replicates are often performed for consistent detection of peptides. We started initially with label-free methods and further coupled pChIP to SILAC method (MS) to obtain both qualitative and quantitative knowledge of proteins dependencies. Coupling pChIP with SILAC and MS reduces the need for replicates as inherently we label the cells with heavy and light stable amino acid isotopes. In addition, a robust pChIP-SILAC ratio is only obtained when the same peptide is detected across both the samples (after mixing). This criterion imposes additional stringency to our robust protein detection, as peptides need to be detected in both heavy and light pChIP samples (a form of replication) and therefore are quantified relative to each other. Another advantage of SILAC method is SILAC-MS software often re-scans the MS spectra to make sure that peptides in both the heavy and light labelled pChIP samples and this is quite critical for low abundant proteins.

To understand the RNAPII regulation and its chromatin proteome, we coupled crosslinked chromatin, pChIP with SILAC-MS to understand and unravel the chromatin proteome. Performing pChIP-SILAC, we realised that several steps in our protocol could significantly impact the final protein (peptide) detection by MS. Additionally, SILAC pChIP experiments designed in complimentary fashion (forward and reverse) yield most robust and consistent protein identification and ratios. The aspects that affect pChIP-SILAC include biological, technical, experimental variability and limitations of current technologies. Some of these include efficiency of SILAC labelling, efficient immunoprecipitation of chromatin complexes, reverse-crosslinking, protein elution and steps leading to MS processing. We measured, optimised and improved these steps leading to detection of large cohorts of proteins;

however further improvement is required and will also be aided by technological advancements in the field of chromatin proteomics.

9.2.2. RNAPII chromatin landscape

RNAPII plays an important role in mES cells and it interacts with a range of proteins during transcription, non-transcriptional processes and importantly cascades a range of additional interaction on chromatin. We performed RNAPII-pChIP (S5p) and were astonished to observe such a dynamic range of RNAPII interacting proteins captured by pChIP. RNAPII-S5p is present at promoters and coding regions of actively transcribing genes and this S5p mark is present in initiating, elongating and terminating RNAPII (Fig. 5.1). By performing RNAPII-S5p pChIP, we capture protein interactions occurring during distinct stages of transcription, non-transcriptional process and importantly mES cell specific processes highlighting the potential of pChIP method. These cohorts include capping enzyme, chromatin remodellers, general transcription factors, TAFs and other proteins associated at promoters or with transcription initiation. We also identified chromatin remodellers, splicing, polyadenylation, co-transcriptional machinery and proteins (phosphatases, export factors) involved in transcription termination. By applying RNAPII-S5p pChIP, we not only identify the known RNAPII interactors but also enrich for novel interactions and important stem cell specific processes. Our laboratory previously identified Polycomb proteins associating with RNAPII-S5p only in mES cells, and our pChIP dataset independently captures and extends our understanding of the RNAPII-Polycomb interplay. In addition, we uncover novel processes associated with RNAPII including cell cycle, DNA replication and metabolism.

Gene expression is mES cells is tightly regulated whereby cells maintain their ability to self-renew and pluripotency while retaining capability to differentiate upon appropriate cues. In mES cells, master regulators (Oct4, Sox2, Nanog, Klf4, etc.) maintain appropriate control of gene expression by restricting or activating the expression of genes. Many master regulators directly or

indirectly interact with RNAPII to mediate this control. The novel association of RNAPII ($S5p^{+}S7p^{-}S2p^{-}$) at important developmental regulator genes along with Polycomb proteins further highlights the importance of RNAPII regulation and its interplay in mES cells. From our single unbiased S5p-pChIP, we uncover many of these associations on chromatin including Polycomb proteins, chromatin remodellers (Hmt, Hat, Hdac etc.), master regulators (Oct4, Sall4) and proteins important for pluripotency (Mediator, TAFs, Esrrb, Paf1). The cohorts of chromatin interactions captured by single pChIP highlights the plethora and vastness of information contained within one pChIP run. The next challenge is to further dissect these interactions asking which proteins associate during active transcription, identifying novel non-transcriptional processes and the protein cohorts involved in the processes. Additionally, understanding whether chromatin components come together at the same time at certain genes or whether these interactions occur at subset of genes governed by temporal regulation would provides us with addition layer of knowledge on RNAPII regulation in stem cells.

9.2.3. Active transcription and RNAPII proteome

During transcription, the Rpb1-CTD of RNAPII undergoes distinct post-translational modifications on serine residues at position 5 (S5p), 7 (S7p) and 2 (S2p) that demarcate transcription initiation, transition to elongation and transcription elongation, respectively. Genome-wide ChIP-Seq distributions of RNAPII modifications at actively transcribing genes (Fig. 5.8) clearly highlight the presence of S5p peaking at promoter and also present throughout the body of a gene. S7p follows a similar pattern to S5p, peaking predominantly at the promoter with lower levels than at the promoter at coding regions. S2p levels are enriched at body of genes increasing towards 3' end of genes (Fig. 5.8 and (Brookes *et al.* 2012). The three different RNAPII modifications provide robust markers for distinct transcriptional stages. We applied this knowledge to identify and unravel protein dependencies during the distinct stages of transcription and also identify non-transcriptional processes with dependencies to distinct RNAPII modifications, in particular S5p which we

knew is associated with Polycomb repression in mES cells. To unravel the dependencies in both qualitative and quantitative manners, we required a comprehensive and complementary experimental setup. Our experimental strategy (Fig. 5.2) was designed to unravel the protein dependencies using two unbiased approaches (universe and pairwise; forward and reverse), with several replicates inbuilt within the experimental setup. Additional information associated with affinity and stoichiometry of interaction relative to RNAPII protein was also encoded in our experimental setup. It is pertinent to mention that our biological question is complex and involves three distinct modifications occurring on a single protein at its CTD, which is highly repetitive. Therefore our comprehensive experimental setup was devised to robustly identify proteins and their dependencies to RNAPII modifications.

The universe and pairwise approach experiments were designed to be complimentary to each other and the combinatorial data analysis robustly unravels the dependencies. The simple classification analysis provides an unbiased and effective measure of understanding the protein dependency to RNAPII modifications based solely on directionality of pChIP-SILAC ratios (Figs. 5.12, 5.13 and 5.14). Using just the directionality (not the magnitude) of pChIP-SILAC ratios, we uncover novel protein associations exclusively with RNAPII-S5p including Polycomb proteins, DNA replication and cohort of chromatin remodellers. Not surprisingly, we observe that most proteins identified are associated with a combination of RNAPII modifications and are involved in transcription including co-transcriptional processing, splicing, polyadenylation and termination. Our simple classification not only robustly identifies the known interactions but also highlights novel protein components that co-associate with RNAPII-occupied chromatin during transcription. The simple classification does not inform well about proteins that have pChIP-SILAC ratios close to Rpb1 ratio (example Paf1 or histones; Fig. 5.17). These proteins being so close to Rpb1 ratios are affected significantly by normalization and therefore their pChIP dependencies are not simply inferred from directionality of ratios. In addition, information pertaining to stoichiometry

or affinity of interaction relative to Rpb1 is not utilised by simple classification and therefore the need for a systems biology approach is apparent to extract information from our comprehensive experiment.

Clustering methods are conventionally used partition datasets whereby groups of objects that are more similar to each other are distinguished from groups of objects that are dissimilar. Clustering is a type of explorative data mining that relies on discontinuity of objects and the partitioning is based on an iterative process of finding and grouping discontinuity. Our comprehensive RNAPII pChIP experiment was designed to unravel the protein dependencies to RNAPII modifications (S5p, S7p and S2p) and the simple analyses showed a continuum of protein association during transcription process. To attempt to unravel in more depth the information encoded in the pChIP-SILAC ratios, we applied a novel systems biology approach (Borislav Vangelov and Mauricio Barahona) including novel clustering and network algorithm that detects patterns within a continuous dataset. We applied conventional PCA (Principle Component Analysis) and other '*à la* PCA' methods; as expected from the complexity of our datasets that study different variants of the same protein (RNAPII) that co-exist in a continuous process (transcription), these methods did not partition the dataset robustly (data not shown). Another incentive in applying our systems approach to pChIP-SILAC ratios is that machine learning methods can detect and unravel patters inherently embedded in the pChIP-SILAC ratios, making them completely unbiased and free of any literary influence. This aspect is especially critical as information on directionality, stoichiometry and affinity of interactions relative to RNAPII is clearly uncovered in the pChIP network (Fig. 6.5). The identification of all detected RNAPII subunits within same cluster and in direct connections highlights the sensitivity and specificity of method to detected and cluster proteins. In addition, we not only robustly identify the protein association as discovered by simple classification (Fig. 5.12) but also partition, in unexpected ways, sub-clusters within the S5p-only proteins and common proteins. Remarkably, we uncover novel components involved in transcription process including

detection of U1 spliceosome subunits, metabolic proteins associated with both S5p&S2p and ribosomal proteins that further suggest the idea of transcription-coupled nuclear translation (Iborra *et al.* 2004; David *et al.* 2012). We also observe reduced detection of ribosomal proteins after treatment with Flavopiridol (transcriptional elongation block) in preliminary RNAPII-S5p pChIP experiments (data not shown). The clustering method detects and performs a gradient separation of the pChIP proteins that range from high levels of S5p only (Clusters 1-3; no S7p or S2p) to proteins with high levels of both S5p and S2p (Cluster 8). It is worthwhile to note that two sub-clusters were obtained enriched for both S5p&S2p, containing splicing factors, mRNA processing factors and components of co-transcription machinery. We believe that these sub-partitioning (cluster 7 and 8) are due to stoichiometry differences and abundance of export factors in cluster 8.

The quantitative information from RNAPII-pChIP also includes protein information and PTM's of histones and other chromatin-bound proteins, such as Rpb1-CTD itself. This is an avenue I hope to explore and further ask if existing and novel PTM can be identified in Rpb1-CTD. Further to explore histones, PTM's on histones and their respective dependencies to RNAPII modifications and across genome. These analyses require optimising, processing and running MS to specifically identify PTM's and further re-analysing MS spectra by different software's to proof-read the modifications.

9.2.4. RNAPII-S5p and novel protein associations.

Analysing our comprehensive experimental setup by either the simple classification or the systems biology approach, we uncover novel proteins associations exclusively with RNAPII-S5p (no S7p or S2p). These interactions are identified due to directionality (enrichment) of pChIP-SILAC ratios, affinity of interaction relative to Rpb1 or stoichiometry of association relative to Rpb1. The S5p-only interactions on chromatin mediate different important biological processes including many stem cell specific processes. The clustering and network sensitively detects further patterns within pChIP-SILAC ratios and we

uncovered three different clusters enriched for RNAPII-S5p only. Reassuringly, we identify Polycomb proteins and DNA replication in Cluster 2 along with many novel chromatin remodellers and silencing proteins. Intriguingly, we also observe some metabolic proteins and isomerases associating with RNAPII-S5p (Cluster 3). Most surprisingly, the clustering identified a highly robust core S5p partition (Cluster 1, Fig. 6.9) consisting of 16 proteins that would not have been detected as important RNAPII-S5p associated proteins by conventional methods. Cluster 1 includes stem cell specific proteins (including Utf1) that are down regulated upon early differentiation to neuronal or cardiac precursors (Carmelo Ferrai, our laboratory; data not shown). Most remarkably, these proteins are robustly detected in pChIP experiment and are relatively abundant in stem cells. Using the pChIP network, we can now zoom in, explore connections and design robust hypothesis-driven experiments to unravel the mechanistic association. We identify Polycomb proteins linked with Ogt and replication proteins (Cluster 2; Fig. 6.8) and now we can ask if all three protein complexes occur at the same time on distinct cohorts of genes or whether one complex interacts with other complexes distinctly and on different genes. Utf1 is a stem cell specific transcription factor and, in our pChIP analyses, it is strongly associated with RNAPII-S5p. This knowledge allows us to identify genomic regions where Utf1 and RNAPII-S5p co-associate and also the other protein cohorts that bind to both these proteins. The association of many silencing factors (clusters 1, 2 and 3) with RNAPII-S5p hints at the complex regulation of RNAPII and suggests that RNAPII-S5p as a wider chromatin regulator that may not only be restricted to actively transcribing genes, Polycomb genes but could also be associated with heterochromatin marks (H3K9me3), tandem repeats (satellites) and interspersed repeats (LINE, SINE) and cascading a local regulation at these regions.

I have demonstrated that by coupling pChIP and SILAC-MS on crosslinked-chromatin, we can unravel and dissect RNAPII chromatin proteome. Performing native pChIP and gradient pChIP, I hope to push the technological

envelope and apply these methods for highly specific questions, *i.e.* protein cohorts that coexist with 'RNAPII-S5p and Polycomb', 'RNAPII-S5p and replication'. Comparing published datasets (Chapter 7) with pChIP-proteins and identifying the type of interaction highlights the versatility and applicability of our approach. These dataset comparisons provide a cost effective bioinformatic way to identify candidates and design appropriate experiments.

9.3. Future research directions

In this thesis, I have focussed on RNAPII chromatin proteome and regulation in mES cells; uncovering novel association of RNAPII-S5p with important biological processes and in addition identifying additional components of transcription machinery. Applying pChIP, we have identified several mES cell-specific proteins involved in transcription regulation and non-transcriptional processes. One of the future aims is to differentiate mES cells to neuronal (or myogenic) differentiation and perform pChIP on differentiated cells. This systems biology comparison would allow us to understand the differences in transcriptional machinery between the cell types, in the pluripotent and post-replicated states, to further unravel regulatory changes in chromatin processes and gene regulation upon lineage. We can further ask whether specific chromatin factors (TFs, chromatin remodellers) regulate the chromatin landscape upon lineage specification or are replaced by neuronal or myogenic-specific chromatin factors. The most important aim would be unravel and understand if the same or different S5p-only processes associate upon lineage specification, mapping genome-wide S5p occupancy and identifying the chromatin proteome by pChIP.

The Rpb1-CTD is highly repetitive and each residue can be subjected to a range of PTM's. Using the Native, gradient RNAPII-pChIP and optimised MS conditions (different peptide cleavers, MS run time, PTM-specific MS analysis), I would hope to detect novel modifications on the Rpb1-CTD. From our native pChIP, we already identify major classes of enzymes (kinases, phosphatases, acetyl-transferases, etc.) and exploring these proteins would

highlight potential candidates that modify Rpb1-CTD among other targets. To further obtain insights on the interplay between DNA replication and RNAPII-S5p in mES cells, we can perform cell synchronisation either using drugs to arrest cell cycle stages (Azuara *et al.* 2006) or by elutriation (Banfalvi 2011). Using the different cell cycle fractions, we can ask whether the RNAPII-S5p and DNA replication association is restricted to specific cell cycle stages and its proteome composition by pChIP. We can further extend this analysis and ask if interesting proteins and their association are cell cycle regulated. With native and gradient chromatin, we have a low complexity, highly concentrated sample that allows us to sensitively capture only the strongest interaction on chromatin (or nucleo-histone fractions). I aim to perform DNA-ChIP and pChIP against RNAPII modifications and Polycomb proteins in these low complexity samples to identify candidate proteins that mediate or maintain the RNAPII-Polycomb interplay. Additionally using Polycomb knockout cell lines ($Ezh2^{-/-}$, $Ring1b^{-/-}$) I would be able to test the functional consequence of the interactions. Finally, pChIP can be applied to study any chromatin process of interest by choosing appropriate combinations of antibodies that help contrast different chromatin states.

10. References

Aebersold, R and Mann, M (2003). Mass spectrometry-based proteomics. Nature **422**, 198-207.

Altelaar, AF, Munoz, J and Heck, AJ (2013). Next-generation proteomics: Towards an integrative view of proteome dynamics. Nature reviews. Genetics **14**, 35-48.

Ang, YS, Tsai, SY, Lee, DF, Monk, J, Su, J, Ratnakumar, K, Ding, J, Ge, Y, Darr, H, Chang, B, Wang, J, Rendl, M, Bernstein, E, Schaniel, C and Lemischka, IR (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell **145**, 183-197.

Azuara, V, Perry, P, Sauer, S, Spivakov, M, Jorgensen, HF, John, RM, Gouti, M, Casanova, M, Warnes, G, Merkenschlager, M and Fisher, AG (2006). Chromatin signatures of pluripotent cell lines. Nature cell biology **8**, 532-538.

Baltz, AG, Munschauer, M, Schwanhausser, B, Vasile, A, Murakawa, Y, Schueler, M, Youngs, N, Penfold-Brown, D, Drew, K, Milek, M, Wyler, E, Bonneau, R, Selbach, M, Dieterich, C and Landthaler, M (2012). The mrna-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell **46**, 674-690.

Banfalvi, G (2011). Synchronization of mammalian cells and nuclei by centrifugal elutriation. Methods in molecular biology **761**, 25-45.

Bannister, AJ and Kouzarides, T (2011). Regulation of chromatin by histone modifications. Cell research **21**, 381-395.

Bartkowiak, B, Liu, P, Phatnani, HP, Fuda, NJ, Cooper, JJ, Price, DH, Adelman, K, Lis, JT and Greenleaf, AL (2010). Cdk12 is a transcription elongation-associated ctd kinase, the metazoan ortholog of yeast ctk1. Genes & development **24**, 2303-2316.

Bartkowiak, B, Mackellar, AL and Greenleaf, AL (2011). Updating the ctd story: From tail to epic. Genetics research international **2011**, 623718.

Baskaran, R, Chiang, GG, Mysliwiec, T, Kruh, GD and Wang, JY (1997). Tyrosine phosphorylation of rna polymerase ii carboxyl-terminal domain by the abl-related gene product. The Journal of biological chemistry **272**, 18905-18909.

Baskaran, R, Chiang, GG and Wang, JY (1996). Identification of a binding site in c-ab1 tyrosine kinase for the c-terminal repeated domain of rna polymerase ii. Molecular and cellular biology **16**, 3361-3369.

Baskaran, R, Dahmus, ME and Wang, JY (1993). Tyrosine phosphorylation of mammalian rna polymerase ii carboxyl-terminal domain. Proceedings of the National Academy of Sciences of the United States of America **90**, 11167-11171.

Baskaran, R, Escobar, SR and Wang, JY (1999). 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**, 387-396.

Bataille, AR, Jeronimo, C, Jacques, PE, Laramee, L, Fortin, ME, Forest, A, Bergeron, M, Hanes, SD and Robert, F (2012). A universal rna polymerase ii ctd cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. Molecular cell **45**, 158-170.

Bendall, SC, Hughes, C, Stewart, MH, Doble, B, Bhatia, M and Lajoie, GA (2008). Prevention of amino acid conversion in silac experiments with embryonic stem cells. Molecular & cellular proteomics : MCP **7**, 1587-1597.

Bernstein, BE, Mikkelsen, TS, Xie, X, Kamal, M, Huebert, DJ, Cuff, J, Fry, B, Meissner, A, Wernig, M, Plath, K, Jaenisch, R, Wagschal, A, Feil, R, Schreiber, SL and Lander, ES (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell **125**, 315-326.

Billon, N, Jolicoeur, C, Ying, QL, Smith, A and Raff, M (2002). Normal timing of oligodendrocyte development from genetically engineered, lineage-selectable mouse es cells. J Cell Sci **115**, 3657-3665.

Black, JC, Choi, JE, Lombardo, SR and Carey, M (2006). A mechanism for coordinating chromatin modification and preinitiation complex assembly. Molecular cell **23**, 809-818.

Blagoev, B and Mann, M (2006). Quantitative proteomics to study mitogen-activated protein kinases. Methods **40**, 243-250.

Boeing, S, Rigault, C, Heidemann, M, Eick, D and Meisterernst, M (2010). Rna polymerase ii c-terminal heptarepeat domain ser-7 phosphorylation is established in a mediator-dependent fashion. The Journal of biological chemistry **285**, 188-196.

Branco, MR and Pombo, A (2007). Chromosome organization: New facts, new models. Trends in cell biology **17**, 127-134.

Breiling, A, Bonte, E, Ferrari, S, Becker, PB and Paro, R (1999). The drosophila polycomb protein interacts with nucleosomal core particles in vitro via its repression domain. Molecular and cellular biology **19**, 8451-8460.

Brinster, RL and Harstad, H (1977). Energy metabolism in primordial germ cells of the mouse. Experimental cell research **109**, 111-117.

Brogna, S, Sato, TA and Rosbash, M (2002). Ribosome components are associated with sites of transcription. Molecular cell **10**, 93-104.

Brookes, E, de Santiago, I, Hebenstreit, D, Morris, KJ, Carroll, T, Xie, SQ, Stock, JK, Heidemann, M, Eick, D, Nozaki, N, Kimura, H, Ragoussis, J, Teichmann, SA and Pombo, A (2012). Polycomb associates genome-wide with a specific rna polymerase ii variant, and regulates metabolic genes in escs. Cell stem cell **10**, 157-170.

Brookes, E and Pombo, A (2009a). Modifications of rna polymerase ii are pivotal in regulating gene expression states. EMBO Rep **10**, 1213-1219.

Brookes, E and Pombo, A (2009b). Modifications of rna polymerase ii are pivotal in regulating gene expression states. EMBO reports **10**, 1213-1219.

Buratowski, S (2003). The ctd code. Nature structural biology **10**, 679-680.
Buratowski, S (2009). Progression through the rna polymerase ii ctd cycle. Molecular cell **36**, 541-546.
Butler, JE and Kadonaga, JT (2002). The rna polymerase ii core promoter: A key component in the regulation of gene expression. Genes & development **16**, 2583-2592.
Cao, R, Wang, L, Wang, H, Xia, L, Erdjument-Bromage, H, Tempst, P, Jones, RS and Zhang, Y (2002). Role of histone h3 lysine 27 methylation in polycomb-group silencing. Science **298**, 1039-1043.
Carrozza, MJ, Li, B, Florens, L, Suganuma, T, Swanson, SK, Lee, KK, Shia, WJ, Anderson, S, Yates, J, Washburn, MP and Workman, JL (2005). Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription. Cell **123**, 581-592.
Cayrou, C, Gregoire, D, Coulombe, P, Danis, E and Mechali, M (2012). Genome-scale identification of active DNA replication origins. Methods **57**, 158-164.
Chambers, I and Smith, A (2004). Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene **23**, 7150-7160.
Chapman, RD, Heidemann, M, Albert, TK, Mailhammer, R, Flatley, A, Meisterernst, M, Kremmer, E and Eick, D (2007). Transcribing rna polymerase ii is phosphorylated at ctd residue serine-7. Science **318**, 1780-1782.
Chapman, RD, Heidemann, M, Hintermair, C and Eick, D (2008). Molecular evolution of the rna polymerase ii ctd. Trends in genetics : TIG **24**, 289-296.
Chapman, RD, Palancade, B, Lang, A, Bensaude, O and Eick, D (2004). The last ctd repeat of the mammalian rna polymerase ii large subunit is important for its stability. Nucleic acids research **32**, 35-44.
Chen, CH (2008). Review of a current role of mass spectrometry for proteome research. Analytica chimica acta **624**, 16-36.
Chen, CT, Hsu, SH and Wei, YH (2012). Mitochondrial bioenergetic function and metabolic plasticity in stem cell differentiation and cellular reprogramming. Biochimica et biophysica acta **1820**, 571-576.
Chen, LL and Carmichael, GG (2010). Long noncoding rnas in mammalian cells: What, where, and why? Wiley interdisciplinary reviews. RNA **1**, 2-21.
Chen, X, Vega, VB and Ng, HH (2008a). Transcriptional regulatory networks in embryonic stem cells. Cold Spring Harbor symposia on quantitative biology **73**, 203-209.
Chen, X, Xu, H, Yuan, P, Fang, F, Huss, M, Vega, VB, Wong, E, Orlov, YL, Zhang, W, Jiang, J, Loh, YH, Yeo, HC, Yeo, ZX, Narang, V, Govindarajan, KR, Leong, B, Shahab, A, Ruan, Y, Bourque, G, Sung, WK, Clarke, ND, Wei, CL and Ng, HH (2008b). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell **133**, 1106-1117.

Cho, EJ, Kobor, MS, Kim, M, Greenblatt, J and Buratowski, S (2001). Opposing effects of ctk1 kinase and fcp1 phosphatase at ser 2 of the rna polymerase ii c-terminal domain. Genes & development **15**, 3319-3329.

Clemente-Blanco, A, Sen, N, Mayan-Santos, M, Sacristan, MP, Graham, B, Jarmuz, A, Giess, A, Webb, E, Game, L, Eick, D, Bueno, A, Merkenschlager, M and Aragon, L (2011). Cdc14 phosphatase promotes segregation of telomeres through repression of rna polymerase ii transcription. Nature cell biology **13**, 1450-1456.

Comer, FI and Hart, GW (2001). Reciprocity between o-glcnac and o-phosphate on the carboxyl terminal domain of rna polymerase ii. Biochemistry **40**, 7845-7852.

Corden, JL, Cadena, DL, Ahearn, JM, Jr. and Dahmus, ME (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic rna polymerase ii. Proceedings of the National Academy of Sciences of the United States of America **82**, 7934-7938.

Cover, TM (1982). Citation classic - nearest neighbor pattern-classification. Current Contents/Engineering Technology & Applied Sciences, 20-20.

Czudnochowski, N, Bosken, CA and Geyer, M (2012). Serine-7 but not serine-5 phosphorylation primes rna polymerase ii ctd for p-tefb recognition. Nature communications **3**, 842.

Dahlberg, JE, Lund, E and Goodwin, EB (2003). Nuclear translation: What is the evidence? RNA **9**, 1-8.

Dai, H, Ciric, B, Zhang, GX and Rostami, A (2012). Interleukin-10 plays a crucial role in suppression of experimental autoimmune encephalomyelitis by bowman-birk inhibitor. Journal of neuroimmunology **245**, 1-7.

Daniel, T and Carling, D (2002). Functional analysis of mutations in the gamma 2 subunit of amp-activated protein kinase associated with cardiac hypertrophy and wolff-parkinson-white syndrome. J Biol Chem **277**, 51017-51024.

Das, R, Yu, J, Zhang, Z, Gygi, MP, Krainer, AR, Gygi, SP and Reed, R (2007). Sr proteins function in coupling rnap ii transcription to pre-mrna splicing. Molecular cell **26**, 867-881.

Daulny, A, Geng, F, Muratani, M, Geisinger, JM, Salghetti, SE and Tansey, WP (2008). Modulation of rna polymerase ii subunit composition by ubiquitylation. Proceedings of the National Academy of Sciences of the United States of America **105**, 19649-19654.

David, A, Dolan, BP, Hickman, HD, Knowlton, JJ, Clavarino, G, Pierre, P, Bennink, JR and Yewdell, JW (2012). Nuclear translation visualized by ribosome-bound nascent chain puromycylation. The Journal of cell biology **197**, 45-57.

de Jong, L, Grande, MA, Mattern, KA, Schul, W and van Driel, R (1996). Nuclear domains involved in rna synthesis, rna processing, and replication. Critical reviews in eukaryotic gene expression **6**, 215-246.

de Napoles, M, Mermoud, JE, Wakao, R, Tang, YA, Endoh, M, Appanah, R, Nesterova, TB, Silva, J, Otte, AP, Vidal, M, Koseki, H and Brockdorff, N

(2004). Polycomb group proteins ring1a/b link ubiquitylation of histone h2a to heritable gene silencing and x inactivation. Dev Cell **7**, 663-676.

Deato, MD and Tjian, R (2008). An unexpected role of tafs and trfs in skeletal muscle differentiation: Switching core promoter complexes. Cold Spring Harbor symposia on quantitative biology **73**, 217-225.

Dejardin, J and Kingston, RE (2009). Purification of proteins associated with specific genomic loci. Cell **136**, 175-186.

Dellino, GI, Schwartz, YB, Farkas, G, McCabe, D, Elgin, SC and Pirrotta, V (2004). Polycomb silencing blocks transcription initiation. Molecular cell **13**, 887-893.

Delvenne, JC, Yaliraki, SN and Barahona, M (2010). Stability of graph communities across time scales. Proceedings of the National Academy of Sciences of the United States of America **107**, 12755-12760.

Devaiah, BN, Lewis, BA, Cherman, N, Hewitt, MC, Albrecht, BK, Robey, PG, Ozato, K, Sims, RJ, 3rd and Singer, DS (2012). Brd4 is an atypical kinase that phosphorylates serine2 of the rna polymerase ii carboxy-terminal domain. Proceedings of the National Academy of Sciences of the United States of America **109**, 6927-6932.

Dieci, G, Fiorino, G, Castelnuovo, M, Teichmann, M and Pagano, A (2007). The expanding rna polymerase iii transcriptome. Trends in genetics : TIG **23**, 614-622.

Ding, J, Xu, H, Faiola, F, Ma'ayan, A and Wang, J (2012). Oct4 links multiple epigenetic pathways to the pluripotency network. Cell research **22**, 155-167.

Ding, L, Paszkowski-Rogacz, M, Nitzsche, A, Slabicki, MM, Heninger, AK, de Vries, I, Kittler, R, Junqueira, M, Shevchenko, A, Schulz, H, Hubner, N, Doss, MX, Sachinidis, A, Hescheler, J, Iacone, R, Anastassiadis, K, Stewart, AF, Pisabarro, MT, Caldarelli, A, Poser, I, Theis, M and Buchholz, F (2009). A genome-scale rnai screen for oct4 modulators defines a role of the paf1 complex for embryonic stem cell identity. Cell stem cell **4**, 403-415.

Efroni, S, Duttagupta, R, Cheng, J, Dehghani, H, Hoeppner, DJ, Dash, C, Bazett-Jones, DP, Le Grice, S, McKay, RD, Buetow, KH, Gingeras, TR, Misteli, T and Meshorer, E (2008). Global transcription in pluripotent embryonic stem cells. Cell stem cell **2**, 437-447.

Egloff, S and Murphy, S (2008a). Cracking the rna polymerase ii ctd code. Trends in genetics : TIG **24**, 280-288.

Egloff, S and Murphy, S (2008b). Role of the c-terminal domain of rna polymerase ii in expression of small nuclear rna genes. Biochemical Society transactions **36**, 537-539.

Egloff, S, O'Reilly, D, Chapman, RD, Taylor, A, Tanzhaus, K, Pitts, L, Eick, D and Murphy, S (2007). Serine-7 of the rna polymerase ii ctd is specifically required for snrna gene expression. Science **318**, 1777-1779.

Evans, MJ and Kaufman, MH (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature **292**, 154-156.

Fabrega, C, Shen, V, Shuman, S and Lima, CD (2003). Structure of an mrna capping enzyme bound to the phosphorylated carboxy-terminal domain of rna polymerase ii. Molecular cell **11**, 1549-1561.

Fischle, W, Wang, Y and Allis, CD (2003). Histone and chromatin cross-talk. Current opinion in cell biology **15**, 172-183.

Follmar, KE, Decroos, FC, Prichard, HL, Wang, HT, Erdmann, D and Olbrich, KC (2006). Effects of glutamine, glucose, and oxygen concentration on the metabolism and proliferation of rabbit adipose-derived stem cells. Tissue engineering **12**, 3525-3533.

Follmer, NE, Wani, AH and Francis, NJ (2012). A polycomb group protein is retained at specific sites on chromatin in mitosis. PLoS genetics **8**, e1003135.

Fong, N, Bird, G, Vigneron, M and Bentley, DL (2003). A 10 residue motif at the c-terminus of the rna pol ii ctd is required for transcription, splicing and 3' end processing. The EMBO journal **22**, 4274-4282.

Fong, YW, Cattoglio, C, Yamaguchi, T and Tjian, R (2012). Transcriptional regulation by coactivators in embryonic stem cells. Trends in cell biology **22**, 292-298.

Fong, YW, Inouye, C, Yamaguchi, T, Cattoglio, C, Grubisic, I and Tjian, R (2011). A DNA repair complex functions as an oct4/sox2 coactivator in embryonic stem cells. Cell **147**, 120-131.

Fuda, NJ, Ardehali, MB and Lis, JT (2009). Defining mechanisms that regulate rna polymerase ii transcription in vivo. Nature **461**, 186-192.

Fukushima, A, Okuda, A, Nishimoto, M, Seki, N, Hori, TA and Muramatsu, M (1998). Characterization of functional domains of an embryonic stem cell coactivator utf1 which are conserved and essential for potentiation of atf-2 activity. The Journal of biological chemistry **273**, 25840-25849.

Galas, DJ and Schmitz, A (1978). Dnase footprinting: A simple method for the detection of protein-DNA binding specificity. Nucleic acids research **5**, 3157-3170.

Galbraith, MD, Donner, AJ and Espinosa, JM (2010). Cdk8: A positive regulator of transcription. Transcription **1**, 4-12.

Garner, MM and Revzin, A (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: Application to components of the escherichia coli lactose operon regulatory system. Nucleic acids research **9**, 3047-3060.

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

Gassmann, M, Fandrey, J, Bichet, S, Wartenberg, M, Marti, HH, Bauer, C, Wenger, RH and Acker, H (1996). Oxygen supply and oxygen-dependent gene expression in differentiating embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America **93**, 2867-2872.

Gesslbauer, B, Krenn, E, Zenzmaier, C, Preisegger, KH and Kungl, AJ (2006). Lessons from the stem cell proteome. Current stem cell research & therapy **1**, 395-409.

Ghazy, MA, He, X, Singh, BN, Hampsey, M and Moore, C (2009). The essential n terminus of the pta1 scaffold protein is required for snorna transcription termination and ssu72 function but is dispensable for pre-mrna 3'-end processing. Molecular and cellular biology **29**, 2296-2307.

Ghosh, A, Shuman, S and Lima, CD (2008). The structure of fcp1, an essential rna polymerase ii ctd phosphatase. Molecular cell **32**, 478-490.

Gilbert, DM (2002). Replication timing and transcriptional control: Beyond cause and effect. Current opinion in cell biology **14**, 377-383.

Gilmour, DS and Lis, JT (1984). Detecting protein-DNA interactions in vivo: Distribution of rna polymerase on specific bacterial genes. Proceedings of the National Academy of Sciences of the United States of America **81**, 4275-4279.

Gilmour, DS and Lis, JT (1985). In vivo interactions of rna polymerase ii with genes of drosophila melanogaster. Molecular and cellular biology **5**, 2009-2018.

Goodrich, JA and Tjian, R (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. Nature reviews. Genetics **11**, 549-558.

Govind, CK, Qiu, H, Ginsburg, DS, Ruan, C, Hofmeyer, K, Hu, C, Swaminathan, V, Workman, JL, Li, B and Hinnebusch, AG (2010). Phosphorylated pol ii ctd recruits multiple hdacs, including rpd3c(s), for methylation-dependent deacetylation of orf nucleosomes. Molecular cell **39**, 234-246.

Graumann, J, Hubner, NC, Kim, JB, Ko, K, Moser, M, Kumar, C, Cox, J, Scholer, H and Mann, M (2008). 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 : MCP **7**, 672-683.

Guillot, PV, Xie, SQ, Hollinshead, M and Pombo, A (2004). Fixation-induced redistribution of hyperphosphorylated rna polymerase ii in the nucleus of human cells. Exp Cell Res **295**, 460-468.

Hafner, M, Landthaler, M, Burger, L, Khorshid, M, Hausser, J, Berninger, P, Rothballer, A, Ascano, M, Jr., Jungkamp, AC, Munschauer, M, Ulrich, A, Wardle, GS, Dewell, S, Zavolan, M and Tuschl, T (2010). Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. Cell **141**, 129-141.

Hajheidari, M, Farrona, S, Huettel, B, Koncz, Z and Koncz, C (2012). Cdkf;1 and cdkd protein kinases regulate phosphorylation of serine residues in the c-terminal domain of arabidopsis rna polymerase ii. The Plant cell **24**, 1626-1642.

Hassan, AB, Errington, RJ, White, NS, Jackson, DA and Cook, PR (1994). Replication and transcription sites are colocalized in human cells. Journal of cell science **107 (Pt 2)**, 425-434.

Hebbes, TR, Thorne, AW and Crane-Robinson, C (1988). A direct link between core histone acetylation and transcriptionally active chromatin. The EMBO journal **7**, 1395-1402.

Heidemann, M, Hintermair, C, Voss, K and Eick, D (2012). Dynamic phosphorylation patterns of rna polymerase ii ctd during transcription. Biochimica et biophysica acta.

Heidemann, M, Hintermair, C, Voss, K and Eick, D (2013). Dynamic phosphorylation patterns of rna polymerase ii ctd during transcription. Biochimica et biophysica acta **1829**, 55-62.

Hengartner, CJ, Myer, VE, Liao, SM, Wilson, CJ, Koh, SS and Young, RA (1998). Temporal regulation of rna polymerase ii by srb10 and kin28 cyclin-dependent kinases. Molecular cell **2**, 43-53.

Henikoff, S, Henikoff, JG, Sakai, A, Loeb, GB and Ahmad, K (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome research **19**, 460-469.

Hintermair, C, Heidemann, M, Koch, F, Descostes, N, Gut, M, Gut, I, Fenouil, R, Ferrier, P, Flatley, A, Kremmer, E, Chapman, RD, Andrau, JC and Eick, D (2012). Threonine-4 of mammalian rna polymerase ii ctd is targeted by polo-like kinase 3 and required for transcriptional elongation. The EMBO journal **31**, 2784-2797.

Hiratani, I, Ryba, T, Itoh, M, Yokochi, T, Schwaiger, M, Chang, CW, Lyou, Y, Townes, TM, Schubeler, D and Gilbert, DM (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS biology **6**, e245.

Hiratani, I, Takebayashi, S, Lu, J and Gilbert, DM (2009). Replication timing and transcriptional control: Beyond cause and effect--part ii. Current opinion in genetics & development **19**, 142-149.

Ho, L, Jothi, R, Ronan, JL, Cui, K, Zhao, K and Crabtree, GR (2009). An embryonic stem cell chromatin remodeling complex, esbaf, is an essential component of the core pluripotency transcriptional network. Proceedings of the National Academy of Sciences of the United States of America **106**, 5187-5191.

Hoffmeyer, K, Raggioli, A, Rudloff, S, Anton, R, Hierholzer, A, Del Valle, I, Hein, K, Vogt, R and Kemler, R (2012). Wnt/beta-catenin signaling regulates telomerase in stem cells and cancer cells. Science **336**, 1549-1554.

Hoshino, A and Fujii, H (2009). Insertional chromatin immunoprecipitation: A method for isolating specific genomic regions. Journal of bioscience and bioengineering **108**, 446-449.

Hsin, JP and Manley, JL (2012). The rna polymerase ii ctd coordinates transcription and rna processing. Genes & development **26**, 2119-2137.

Hsin, JP, Sheth, A and Manley, JL (2011). Rnap ii ctd phosphorylated on threonine-4 is required for histone mrna 3' end processing. Science **334**, 683-686.

Iborra, FJ, Jackson, DA and Cook, PR (2001). Coupled transcription and translation within nuclei of mammalian cells. Science **293**, 1139-1142.

Iborra, FJ, Jackson, DA and Cook, PR (2004). The case for nuclear translation. Journal of cell science **117**, 5713-5720.

Ip, YT, Jackson, V, Meier, J and Chalkley, R (1988). The separation of transcriptionally engaged genes. The Journal of biological chemistry **263**, 14044-14052.

Izaurralde, E, Lewis, J, McGuigan, C, Jankowska, M, Darzynkiewicz, E and Mattaj, IW (1994). A nuclear cap binding protein complex involved in pre-mrna splicing. Cell **78**, 657-668.

Jackson, DA and Pombo, A (1998). Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of s phase in human cells. The Journal of cell biology **140**, 1285-1295.

Jaenisch, R and Young, R (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell **132**, 567-582.

Jang, MK, Mochizuki, K, Zhou, M, Jeong, HS, Brady, JN and Ozato, K (2005). The bromodomain protein brd4 is a positive regulatory component of p-tefb and stimulates rna polymerase ii-dependent transcription. Molecular cell **19**, 523-534.

Jenuwein, T and Allis, CD (2001). Translating the histone code. Science **293**, 1074-1080.

Jeronimo, C, Forget, D, Bouchard, A, Li, Q, Chua, G, Poitras, C, Therien, C, Bergeron, D, Bourassa, S, Greenblatt, J, Chabot, B, Poirier, GG, Hughes, TR, Blanchette, M, Price, DH and Coulombe, B (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7sk capping enzyme. Molecular cell **27**, 262-274.

Jia, J, Zheng, X, Hu, G, Cui, K, Zhang, J, Zhang, A, Jiang, H, Lu, B, Yates, J, 3rd, Liu, C, Zhao, K and Zheng, Y (2012). Regulation of pluripotency and self- renewal of escs through epigenetic- threshold modulation and mrna pruning. Cell **151**, 576-589.

Kagey, MH, Newman, JJ, Bilodeau, S, Zhan, Y, Orlando, DA, van Berkum, NL, Ebmeier, CC, Goossens, J, Rahl, PB, Levine, SS, Taatjes, DJ, Dekker, J and Young, RA (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature **467**, 430-435.

Keene, JD, Komisarow, JM and Friedersdorf, MB (2006). Rip-chip: The isolation and identification of mrnas, microrna and protein components of ribonucleoprotein complexes from cell extracts. Nature protocols **1**, 302-307.

Kelly, WG, Dahmus, ME and Hart, GW (1993). Rna polymerase ii is a glycoprotein. Modification of the cooh-terminal domain by o-glcnac. The Journal of biological chemistry **268**, 10416-10424.

Keogh, MC, Kurdistani, SK, Morris, SA, Ahn, SH, Podolny, V, Collins, SR, Schuldiner, M, Chin, K, Punna, T, Thompson, NJ, Boone, C, Emili, A, Weissman, JS, Hughes, TR, Strahl, BD, Grunstein, M, Greenblatt, JF, Buratowski, S and Krogan, NJ (2005). Cotranscriptional set2 methylation of histone h3 lysine 36 recruits a repressive rpd3 complex. Cell **123**, 593-605.

Kim, TS, Liu, CL, Yassour, M, Holik, J, Friedman, N, Buratowski, S and Rando, OJ (2010). Rna polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast. Genome biology **11**, R75.

Kim, YK, Bourgeois, CF, Isel, C, Churcher, MJ and Karn, J (2002). Phosphorylation of the rna polymerase ii carboxyl-terminal domain by cdk9 is directly responsible for human immunodeficiency virus type 1 tat-activated transcriptional elongation. Molecular and cellular biology **22**, 4622-4637.

Kizer, KO, Phatnani, HP, Shibata, Y, Hall, H, Greenleaf, AL and Strahl, BD (2005). A novel domain in set2 mediates rna polymerase ii interaction and couples histone h3 k36 methylation with transcript elongation. Molecular and cellular biology **25**, 3305-3316.

Komarnitsky, P, Cho, EJ and Buratowski, S (2000). Different phosphorylated forms of rna polymerase ii and associated mrna processing factors during transcription. Genes & development **14**, 2452-2460.

Konig, J, Zarnack, K, Luscombe, NM and Ule, J (2011). Protein-rna interactions: New genomic technologies and perspectives. Nature reviews. Genetics **13**, 77-83.

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

Krishnamurthy, S, Ghazy, MA, Moore, C and Hampsey, M (2009). Functional interaction of the ess1 prolyl isomerase with components of the rna polymerase ii initiation and termination machineries. Molecular and cellular biology **29**, 2925-2934.

Ku, M, Koche, RP, Rheinbay, E, Mendenhall, EM, Endoh, M, Mikkelsen, TS, Presser, A, Nusbaum, C, Xie, X, Chi, AS, Adli, M, Kasif, S, Ptaszek, LM, Cowan, CA, Lander, ES, Koseki, H and Bernstein, BE (2008). Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. PLoS genetics **4**, e1000242.

Kuzmichev, A, Nishioka, K, Erdjument-Bromage, H, Tempst, P and Reinberg, D (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the enhancer of zeste protein. Genes Dev **16**, 2893-2905.

Lambert, JP, Baetz, K and Figeys, D (2010). Of proteins and DNA--proteomic role in the field of chromatin research. Molecular bioSystems **6**, 30-37.

Lambert, JP, Mitchell, L, Rudner, A, Baetz, K and Figeys, D (2009). A novel proteomics approach for the discovery of chromatin-associated protein networks. Molecular & cellular proteomics : MCP **8**, 870-882.

Lee, JT (2012). Epigenetic regulation by long noncoding rnas. Science **338**, 1435-1439.

Leeb, M and Wutz, A (2012). Establishment of epigenetic patterns in development. Chromosoma **121**, 251-262.

Li, B, Carey, M and Workman, JL (2007a). The role of chromatin during transcription. Cell **128**, 707-719.

Li, H, Zhang, Z, Wang, B, Zhang, J, Zhao, Y and Jin, Y (2007b). Wwp2-mediated ubiquitination of the rna polymerase ii large subunit in mouse

embryonic pluripotent stem cells. Molecular and cellular biology **27**, 5296-5305.

Li, M, Liu, GH and Izpisua Belmonte, JC (2012). Navigating the epigenetic landscape of pluripotent stem cells. Nature reviews. Molecular cell biology **13**, 524-535.

Li, M, Phatnani, HP, Guan, Z, Sage, H, Greenleaf, AL and Zhou, P (2005). Solution structure of the set2-rpb1 interacting domain of human set2 and its interaction with the hyperphosphorylated c-terminal domain of rpb1. Proceedings of the National Academy of Sciences of the United States of America **102**, 17636-17641.

Liu, Z, Scannell, DR, Eisen, MB and Tjian, R (2011). Control of embryonic stem cell lineage commitment by core promoter factor, taf3. Cell **146**, 720-731.

Luis, NM, Morey, L, Di Croce, L and Benitah, SA (2012). Polycomb in stem cells: Prc1 branches out. Cell stem cell **11**, 16-21.

Lund, AH and van Lohuizen, M (2004). Polycomb complexes and silencing mechanisms. Current opinion in cell biology **16**, 239-246.

Mann, M (2006). Functional and quantitative proteomics using silac. Nature reviews. Molecular cell biology **7**, 952-958.

Martin, GR (1981). 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 of the United States of America-Biological Sciences **78**, 7634-7638.

Medlin, J, Scurry, A, Taylor, A, Zhang, F, Peterlin, BM and Murphy, S (2005). P-tefb is not an essential elongation factor for the intronless human u2 snrna and histone h2b genes. The EMBO journal **24**, 4154-4165.

Meininghaus, M, Chapman, RD, Horndasch, M and Eick, D (2000). Conditional expression of rna polymerase ii in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription. The Journal of biological chemistry **275**, 24375-24382.

Melcer, S and Meshorer, E (2010). Chromatin plasticity in pluripotent cells. Essays in biochemistry **48**, 245-262.

Meshorer, E, Yellajoshula, D, George, E, Scambler, PJ, Brown, DT and Misteli, T (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Developmental cell **10**, 105-116.

Mikkelsen, TS, Ku, M, Jaffe, DB, Issac, B, Lieberman, E, Giannoukos, G, Alvarez, P, Brockman, W, Kim, TK, Koche, RP, Lee, W, Mendenhall, E, O'Donovan, A, Presser, A, Russ, C, Xie, X, Meissner, A, Wernig, M, Jaenisch, R, Nusbaum, C, Lander, ES and Bernstein, BE (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448**, 553-560.

Moller, A, Xie, SQ, Hosp, F, Lang, B, Phatnani, HP, James, S, Ramirez, F, Collin, GB, Naggert, JK, Babu, MM, Greenleaf, AL, Selbach, M and Pombo, A (2012). Proteomic analysis of mitotic rna polymerase ii reveals novel interactors and association with proteins dysfunctional in disease. Molecular & cellular proteomics : MCP **11**, M111 011767.

Morris, DP, Michelotti, GA and Schwinn, DA (2005). Evidence that phosphorylation of the rna polymerase ii carboxyl-terminal repeats is similar in yeast and humans. The Journal of biological chemistry **280**, 31368-31377.

Morris, EJ, Ji, JY, Yang, F, Di Stefano, L, Herr, A, Moon, NS, Kwon, EJ, Haigis, KM, Naar, AM and Dyson, NJ (2008). E2f1 represses beta-catenin transcription and is antagonized by both prb and cdk8. Nature **455**, 552-556.

Mosley, AL, Pattenden, SG, Carey, M, Venkatesh, S, Gilmore, JM, Florens, L, Workman, JL and Washburn, MP (2009). Rtr1 is a ctd phosphatase that regulates rna polymerase ii during the transition from serine 5 to serine 2 phosphorylation. Molecular cell **34**, 168-178.

Myers, SA, Panning, B and Burlingame, AL (2011). Polycomb repressive complex 2 is necessary for the normal site-specific o-glcnac distribution in mouse embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America **108**, 9490-9495.

Nechaev, S and Adelman, K (2011). Pol ii waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochimica et biophysica acta **1809**, 34-45.

Niwa, H, Miyazaki, J and Smith, AG (2000). Quantitative expression of oct-3/4 defines differentiation, dedifferentiation or self-renewal of es cells. Nat Genet **24**, 372-376.

O'Carroll, D, Erhardt, S, Pagani, M, Barton, SC, Surani, MA and Jenuwein, T (2001). The polycomb-group gene ezh2 is required for early mouse development. Molecular and cellular biology **21**, 4330-4336.

O'Neill, LP and Turner, BM (2003). Immunoprecipitation of native chromatin: Nchip. Methods **31**, 76-82.

Ong, SE, Blagoev, B, Kratchmarova, I, Kristensen, DB, Steen, H, Pandey, A and Mann, M (2002). Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics : MCP **1**, 376-386.

Ong, SE and Mann, M (2005). Mass spectrometry-based proteomics turns quantitative. Nature chemical biology **1**, 252-262.

Orkin, SH, Wang, J, Kim, J, Chu, J, Rao, S, Theunissen, TW, Shen, X and Levasseur, DN (2008). The transcriptional network controlling pluripotency in es cells. Cold Spring Harbor symposia on quantitative biology **73**, 195-202.

Orlando, V, Strutt, H and Paro, R (1997). Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods **11**, 205-214.

Ouwerkerk, PB and Meijer, AH (2001). Yeast one-hybrid screening for DNA-protein interactions. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] **Chapter 12**, Unit 12 12.

Pagano, A, Castelnuovo, M, Tortelli, F, Ferrari, R, Dieci, G and Cancedda, R (2007). New small nuclear rna gene-like transcriptional units as sources of regulatory transcripts. PLoS genetics **3**, e1.

Pan, G, Tian, S, Nie, J, Yang, C, Ruotti, V, Wei, H, Jonsdottir, GA, Stewart, R and Thomson, JA (2007). Whole-genome analysis of histone h3 lysine

4 and lysine 27 methylation in human embryonic stem cells. Cell stem cell **1**, 299-312.
Phatnani, HP and Greenleaf, AL (2006). Phosphorylation and functions of the rna polymerase ii ctd. Genes & development **20**, 2922-2936.
Phizicky, EM and Fields, S (1995). Protein-protein interactions: Methods for detection and analysis. Microbiological reviews **59**, 94-123.
Pirngruber, J, Shchebet, A, Schreiber, L, Shema, E, Minsky, N, Chapman, RD, Eick, D, Aylon, Y, Oren, M and Johnsen, SA (2009). Cdk9 directs h2b monoubiquitination and controls replication-dependent histone mrna 3'-end processing. EMBO reports **10**, 894-900.
Pirrotta, V and Li, HB (2012). A view of nuclear polycomb bodies. Current opinion in genetics & development **22**, 101-109.
Prieto, C and De Las Rivas, J (2006). Apid: Agile protein interaction dataanalyzer. Nucleic acids research **34**, W298-302.
Proudfoot, NJ, Furger, A and Dye, MJ (2002). Integrating mrna processing with transcription. Cell **108**, 501-512.
Prudhomme, W, Daley, GQ, Zandstra, P and Lauffenburger, DA (2004). Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. Proceedings of the National Academy of Sciences of the United States of America **101**, 2900-2905.
Rahl, PB, Lin, CY, Seila, AC, Flynn, RA, McCuine, S, Burge, CB, Sharp, PA and Young, RA (2010). C-myc regulates transcriptional pause release. Cell **141**, 432-445.
Ranuncolo, SM, Ghosh, S, Hanover, JA, Hart, GW and Lewis, BA (2012). Evidence of the involvement of o-glcnac-modified human rna polymerase ii ctd in transcription in vitro and in vivo. The Journal of biological chemistry **287**, 23549-23561.
Reneker, JS and Brotherton, TW (1991). Discrete regions of the avian beta-globin gene cluster have tissue-specific hypersensitivity to cleavage by sonication in nuclei. Nucleic acids research **19**, 4739-4745.
Richard, P and Manley, JL (2009). Transcription termination by nuclear rna polymerases. Genes & development **23**, 1247-1269.
Ringrose, L and Paro, R (2007). Polycomb/trithorax response elements and epigenetic memory of cell identity. Development **134**, 223-232.
Roeder, RG (2005). Transcriptional regulation and the role of diverse coactivators in animal cells. FEBS letters **579**, 909-915.
Schaub, MT, Delvenne, JC, Yaliraki, SN and Barahona, M (2012). Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. PloS one **7**, e32210.
Schwartz, YB, Kahn, TG, Dellino, GI and Pirrotta, V (2004). Polycomb silencing mechanisms in drosophila. Cold Spring Harbor symposia on quantitative biology **69**, 301-308.
Schwartz, YB, Kahn, TG and Pirrotta, V (2005). Characteristic low density and shear sensitivity of cross-linked chromatin containing polycomb complexes. Molecular and cellular biology **25**, 432-439.
Segal, E and Widom, J (2009). What controls nucleosome positions? Trends in genetics : TIG **25**, 335-343.

Sequeira-Mendes, J, Diaz-Uriarte, R, Apedaile, A, Huntley, D, Brockdorff, N and Gomez, M (2009). Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS genetics **5**, e1000446.

Silva, J, Nichols, J, Theunissen, TW, Guo, G, van Oosten, AL, Barrandon, O, Wray, J, Yamanaka, S, Chambers, I and Smith, A (2009). Nanog is the gateway to the pluripotent ground state. Cell **138**, 722-737.

Sims, RJ, 3rd and Reinberg, D (2009). Processing the h3k36me3 signature. Nature genetics **41**, 270-271.

Sims, RJ, 3rd, Rojas, LA, Beck, D, Bonasio, R, Schuller, R, Drury, WJ, 3rd, Eick, D and Reinberg, D (2011). The c-terminal domain of rna polymerase ii is modified by site-specific methylation. Science **332**, 99-103.

Smith, AG, Heath, JK, Donaldson, DD, Wong, GG, Moreau, J, Stahl, M and Rogers, D (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature **336**, 688-690.

Smith, TA and Hooper, ML (1983). Medium conditioned by feeder cells inhibits the differentiation of embryonal carcinoma cultures. Experimental cell research **145**, 458-462.

Solomon, MJ, Larsen, PL and Varshavsky, A (1988). Mapping protein-DNA interactions in vivo with formaldehyde: Evidence that histone h4 is retained on a highly transcribed gene. Cell **53**, 937-947.

Stock, JK, Giadrossi, S, Casanova, M, Brookes, E, Vidal, M, Koseki, H, Brockdorff, N, Fisher, AG and Pombo, A (2007a). Ring1-mediated ubiquitination of h2a restrains poised rna polymerase ii at bivalent genes in mouse es cells. Nat Cell Biol **9**, 1428-1435.

Stock, JK, Giadrossi, S, Casanova, M, Brookes, E, Vidal, M, Koseki, H, Brockdorff, N, Fisher, AG and Pombo, A (2007b). Ring1-mediated ubiquitination of h2a restrains poised rna polymerase ii at bivalent genes in mouse es cells. Nature cell biology **9**, 1428-1435.

Strahl, BD and Allis, CD (2000). The language of covalent histone modifications. Nature **403**, 41-45.

Sury, MD, Chen, JX and Selbach, M (2010). The silac fly allows for accurate protein quantification in vivo. Molecular & cellular proteomics : MCP **9**, 2173-2183.

Szutorisz, H, Canzonetta, C, Georgiou, A, Chow, CM, Tora, L and Dillon, N (2005). Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. Mol Cell Biol **25**, 1804-1820.

Teves, SS and Henikoff, S (2011). Heat shock reduces stalled rna polymerase ii and nucleosome turnover genome-wide. Genes & development **25**, 2387-2397.

Tritschler, F, Braun, JE, Motz, C, Igreja, C, Haas, G, Truffault, V, Izaurralde, E and Weichenrieder, O (2009). Dcp1 forms asymmetric trimers to assemble into active mrna decapping complexes in metazoa. Proceedings of the National Academy of Sciences of the United States of America **106**, 21591-21596.

Udy, GB, Parkes, BD and Wells, DN (1997). Es cell cycle rates affect gene targeting frequencies. Experimental cell research **231**, 296-301.

Uniprot (2012). Reorganizing the protein space at the universal protein resource (uniprot). Nucleic acids research **40**, D71-75.

Unwin, RD, Gaskell, SJ, Evans, CA and Whetton, AD (2003). The potential for proteomic definition of stem cell populations. Experimental hematology **31**, 1147-1159.

Valk-Lingbeek, ME, Bruggeman, SW and van Lohuizen, M (2004). Stem cells and cancer; the polycomb connection. Cell **118**, 409-418.

van den Berg, DL, Snoek, T, Mullin, NP, Yates, A, Bezstarosti, K, Demmers, J, Chambers, I and Poot, RA (2010). An oct4-centered protein interaction network in embryonic stem cells. Cell stem cell **6**, 369-381.

Van Hoof, D, Mummery, CL, Heck, AJ and Krijgsveld, J (2006). Embryonic stem cell proteomics. Expert review of proteomics **3**, 427-437.

Van Hoof, D, Pinkse, MW, Oostwaard, DW, Mummery, CL, Heck, AJ and Krijgsveld, J (2007). An experimental correction for arginine-to-proline conversion artifacts in silac-based quantitative proteomics. Nature methods **4**, 677-678.

van Steensel, B (2011). Chromatin: Constructing the big picture. The EMBO journal **30**, 1885-1895.

van Steensel, B, Delrow, J and Henikoff, S (2001). Chromatin profiling using targeted DNA adenine methyltransferase. Nature genetics **27**, 304-308.

Vaquerizas, JM, Kummerfeld, SK, Teichmann, SA and Luscombe, NM (2009). A census of human transcription factors: Function, expression and evolution. Nature reviews. Genetics **10**, 252-263.

Varelas, X, Sakuma, R, Samavarchi-Tehrani, P, Peerani, R, Rao, BM, Dembowy, J, Yaffe, MB, Zandstra, PW and Wrana, JL (2008). Taz controls smad nucleocytoplasmic shuttling and regulates human embryonic stem-cell self-renewal. Nature cell biology **10**, 837-848.

Venkatesan, K, Rual, JF, Vazquez, A, Stelzl, U, Lemmens, I, Hirozane-Kishikawa, T, Hao, T, Zenkner, M, Xin, X, Goh, KI, Yildirim, MA, Simonis, N, Heinzmann, K, Gebreab, F, Sahalie, JM, Cevik, S, Simon, C, de Smet, AS, Dann, E, Smolyar, A, Vinayagam, A, Yu, H, Szeto, D, Borick, H, Dricot, A, Klitgord, N, Murray, RR, Lin, C, Lalowski, M, Timm, J, Rau, K, Boone, C, Braun, P, Cusick, ME, Roth, FP, Hill, DE, Tavernier, J, Wanker, EE, Barabasi, AL and Vidal, M (2009). An empirical framework for binary interactome mapping. Nature methods **6**, 83-90.

Wade, JT and Struhl, K (2008). The transition from transcriptional initiation to elongation. Current opinion in genetics & development **18**, 130-136.

Walsh, MJ, Hautbergue, GM and Wilson, SA (2010). Structure and function of mrna export adaptors. Biochemical Society transactions **38**, 232-236.

Wang, H, Wang, L, Erdjument-Bromage, H, Vidal, M, Tempst, P, Jones, RS and Zhang, Y (2004). Role of histone h2a ubiquitination in polycomb silencing. Nature **431**, 873-878.

Wang, J and Orkin, SH (2008). A protein roadmap to pluripotency and faithful reprogramming. Cells, tissues, organs **188**, 23-30.

Wang, J, Rao, S, Chu, J, Shen, X, Levasseur, DN, Theunissen, TW and Orkin, SH (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature **444**, 364-368.

Werner-Allen, JW, Lee, CJ, Liu, P, Nicely, NI, Wang, S, Greenleaf, AL and Zhou, P (2011). Cis-proline-mediated ser(p)5 dephosphorylation by the rna polymerase ii c-terminal domain phosphatase ssu72. The Journal of biological chemistry **286**, 5717-5726.

West, S, Proudfoot, NJ and Dye, MJ (2008). Molecular dissection of mammalian rna polymerase ii transcriptional termination. Molecular cell **29**, 600-610.

Williams, RL, Hilton, DJ, Pease, S, Willson, TA, Stewart, CL, Gearing, DP, Wagner, EF, Metcalf, D, Nicola, NA and Gough, NM (1988). Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. Nature **336**, 684-687.

Wu, H, Min, J, Lunin, VV, Antoshenko, T, Dombrovski, L, Zeng, H, Allali-Hassani, A, Campagna-Slater, V, Vedadi, M, Arrowsmith, CH, Plotnikov, AN and Schapira, M (2010). Structural biology of human h3k9 methyltransferases. PloS one **5**, e8570.

Xie, SQ, Martin, S, Guillot, PV, Bentley, DL and Pombo, A (2006). Splicing speckles are not reservoirs of rna polymerase ii, but contain an inactive form, phosphorylated on serine2 residues of the c-terminal domain. Mol Biol Cell **17**, 1723-1733.

Xie, SQ and Pombo, A (2006). Distribution of different phosphorylated forms of rna polymerase ii in relation to cajal and pml bodies in human cells: An ultrastructural study. Histochemistry and cell biology **125**, 21-31.

Yang, W, Xia, Y, Hawke, D, Li, X, Liang, J, Xing, D, Aldape, K, Hunter, T, Alfred Yung, WK and Lu, Z (2012). Pkm2 phosphorylates histone h3 and promotes gene transcription and tumorigenesis. Cell **150**, 685-696.

Yeo, M, Lee, SK, Lee, B, Ruiz, EC, Pfaff, SL and Gill, GN (2005). Small ctd phosphatases function in silencing neuronal gene expression. Science **307**, 596-600.

Ying, QL, Nichols, J, Chambers, I and Smith, A (2003). Bmp induction of id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with stat3. Cell **115**, 281-292.

Young, RA (2011). Control of the embryonic stem cell state. Cell **144**, 940-954.

Zanivan, S, Krueger, M and Mann, M (2012). In vivo quantitative proteomics: The silac mouse. Methods in molecular biology **757**, 435-450.

Zeng, PY, Vakoc, CR, Chen, ZC, Blobel, GA and Berger, SL (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. BioTechniques **41**, 694, 696, 698.

Zhang, DW, Mosley, AL, Ramisetty, SR, Rodriguez-Molina, JB, Washburn, MP and Ansari, AZ (2012a). Ssu72 phosphatase-dependent erasure of phospho-ser7 marks on the rna polymerase ii c-terminal domain is essential for viability and transcription termination. The Journal of biological chemistry **287**, 8541-8551.

Zhang, DW, Rodriguez-Molina, JB, Tietjen, JR, Nemec, CM and Ansari, AZ (2012b). Emerging views on the ctd code. Genetics research international **2012**, 347214.

Zhou, W, Liotta, LA and Petricoin, EF (2012). Cancer metabolism: What we can learn from proteomic analysis by mass spectrometry. Cancer genomics & proteomics **9**, 373-381.