RNA Polymerase II identifies enhancers in different states of activation

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

bzw. Doctor of Philosophy (Ph.D.)

eingereicht an der

Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

M.Sc. Giulia Caglio

Präsidentin

der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

Prof. Dr. rer. nat. Bernhard Grimm

Gutachter/innen

1. Prof. Dr. Ana Pombo
2. Prof. Dr. Leonie Ringrose
3. Prof. Dr. Norbert Huebner

Tag der mündlichen Prüfung: 13.06.2018

Chi mette il piè su l'amorosa pania,
Cerchi ritrarlo, e non v'inveschi l'ale;
Che non è in somma amor, se non insania,
A giudizio de' savi universale:
E se ben come Orlando ognun non smania,
Suo furor mostra a qualch'altro segnale.
E quale è di pazzia segno più espresso
Che, per altri voler, perder se stesso?

L'Orlando Furioso (Canto XXIV, [I])

Let him make haste his feet to disengage,
Nor lime his wings, whom Love has made a prize;
For love, in fine, is nought but phrensied rage,
By universal suffrage of the wise:
And albeit some may show themselves more sage
Than Roland, they nut sin in other guise.
For, what proves folly more than on this shelf,
Thus, for another, to destroy oneself?

L'Orlando Furioso (Canto XXIV, [I])

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015.

Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/ Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Declaration

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Berlin, ……..

……..……..……..……..

Giulia Caglio

Abstract

Enhancers regulate transcription of target genes and gene expression. They act as recruitment sites for multiple transcription factors (TFs) and RNA polymerase II (RNAPII) and favour transcription of target genes through chromatin contacts. RNAPII at enhancer regions transcribes short and mostly non-polyadenylated transcripts, called enhancer RNAs (eRNAs).

The mechanisms of RNAPII recruitment and regulation at enhancers remain ill understood, in particular how signalling through RNAPII modifications may influence chromatin states, looping and gene activation. In this study, I compare enhancer lists defined with different approaches and find that their relation is very complex. However, I find that RNAPII binding co-occurs with TF binding at regulatory regions, independently of the identification approach used. I characterize the state of RNAPII activation at enhancers and its transcriptional activity. I find that RNAPII state reflects enhancer activation state and correlates with different transcriptional outputs. In addition, I demonstrate that extragenic RNAPII is a novel tool to identify regulatory regions. I successfully identified putative regulatory regions in mESC and during neuronal differentiation, with enhancer activity *in vivo*. Extragenic RNAPII regions have specific activation patterns during neuronal differentiation, are finely regulated at the transcriptional level by kinases and transcribe differently mature RNAs.

In conclusion, I establish RNAPII as a tool to identify and characterise regulatory regions in a cell type of interest. With minimal RNAPII datasets it is possible to simultaneously identify regulatory regions, infer their state of activation, and the state of activation of coding gene promoters.

Zusammenfassung

Enhancer regulieren die Transkription ihrer Zielgene und deren Expression. Sie bieten eine Bindestelle für verschiedenste Transkriptionsfaktoren (TF) und RNA Polymerase II (RNAPII) und unterstützen die Gentranskription durch das Zustandekommen von Chromatinkontakten. Zusätzlich transkribiert RNAPII in Enhancer-Regionen kurze, non-polyadenylierte Transkripte, die man Enhancer-RNA (eRNA) nennt. Der Mechanismus der RNAPII-Rekrutierung und –Regulation an Enhancern ist bisher wenig verstanden, insbesondere wie das Vorhandensein von RNAPII-Modifikationen den Chromatinstatus, -faltung sowie die Genaktivierung beeinflusst.

In dieser Arbeit wurden verschiedene Ansätze der Enhancer-Bestimmung miteinander verglichen. Während eine klare Bestimmung des besten Ansatzes sich als komplex erwies, konnte gezeigt werden, dass die Bindung von RNAPII an regulatorische Regionen in Zusammenhang mit TF eine universelle Konstante darstellte. Weiterhin wurden der Status der Enhancer-gekoppelten RNAPII-Aktivierung und deren Transkriptionsaktivität untersucht. Als Hauptergebnis ergab sich, dass der RNAPII-Status mit der Enhancer-Aktivität und daraus folgend mit veränderter Transkriptionsaktivität korreliert ist. Weiterhin konnte gezeigt werden, dass das Vorhandensein extragenischer RNAPII ein neues Werkzeug zur Identifikation von regulatorischen Regionen ist. Erfolgreich konnten regulatorische Regionen in embryonalen Stammzellen der Maus sowie während der neuronalen Differenzierung vorhergesagt und mittels Enhancer-Aktivität *in-vivo* bestätigt werden. Dabei zeigte sich, dass im Laufe der der neuronalen Differenzierung extragenische RNAPII-Bindung spezifische Aktivierungsmuster aufweist: ihr Transkriptionslevel wird durch Kinasen feinmaschig reguliert und es werden verschiedene Formen maturierter RNA erzeugt.

Zusammenfassend konnte RNAPII als Werkzeug zur Identifikation und Charakterisierung regulatorischer Regionen in verschiedenen Zelltypen ausgemacht werden. Selbst mit minimalen RNAPII-Datensätzen ist es möglich, gleichzeitig regulatorische Regionen zu identifizieren als auch ihren eigenen Aktivierungsstatus sowie den ihrer kodierender Genpromotoren zu bestimmen.

Table of Contents

Index Figures

Index Tables

Publications

Caglio G, Torlai Triglia E, Pombo A. 2017. Keep Them Close: PRC2 Poises Enhancer-Promoter Interactions at Anterior Neuronal Genes. *Cell Stem Cell.* 20:573-575

Kollet O, Vagima Y, D'Uva G, Golan K, Canaani J, Itkin T, Gur-Cohen S, Kalinkovich A, **Caglio G**, Medaglia C, Ludin A, Lapid K, Shezen E, Neufeld-Cohen A, Varol D, Chen A, Lapidot T. 2013 Physiologic corticosterone oscillations regulate murine hematopoietic stem/progenitor cell proliferation and CXCL12 expression by bone marrow stromal progenitors. *Leukemia*. 27:2006-15

Lapid K, Itkin T, D'Uva G, Ovadya Y, Ludin A, **Caglio G**, Kalinkovich A, Golan K, Porat Z, Zollo M, Lapidot T. 2013. GSK3β regulates physiological migration of stem/progenitor cells via cytoskeletal rearrangement. *J Clin Investigation.* 123:1705-17

Abbreviations

3C	Chromosome Conformation Capture
3D	three-dimensional
4C	Circularized Chromosome Conformation Capture
5fC	5-formylcytosin
5hmC	5-hydroxymethylcytosine
8WG16	Antibody recognising RNAPII S2 unphosphorylated
Ac	acetylation
BCP	Bayesian Change-Point
bp	base pair
BWA	Burrows-Wheeler Aligner
CAGE	Cap analysis of gene expression
CBP	CREB-binding protein
Cdk	cycline-dependent kinase
ChIAPet	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin-immunoprecipitation
ChIP-Seq	Chromatin-immunoprecipitation followed by high-throughput sequencing
CRISPR-Cas9	CRISPR-associated protein-9 nuclease
CTCF	CCCTC-binding factor
CTD	C-terminal domain
DNA	Deoxyribonucleic Acid
E11.5	Embryonic day 11.5
ENCODE	Encyclopedia of DNA Elements
FISH	Fluorescent in Situ Hybridisation
FP	False Positive

FPKMs	Fragments Per Kilobase of exon per Million reads mapped
GAM	Genome Architecture Mapping
GO	Gene Ontology
GREAT	Genomic Region Erichment of Annotation Tool
GRO-seq	Global run on sequencing
H3K122ac	Histone 3 lysine 122 acetylation
H3K27ac/me3	Histone 3 lysine 27 acetylation/trimethylation
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4me1/3	Histone 3 lysine 4 mono/trimethylation
H3K64ac	Histone 3 lysine 64 acetylation
Hi-C	High-throughput sequencing Chromosome conformation capture
HM	Histone Modifications
IGV	Integrative Genomics Viewer
IP	Immunoprecipitation
kb	kilo bases
KD	Knock Down
KO	Knock Out
LCR	Locus Control Region
Lif	Leukaemia inhibitory factor
lncRNAs	long non-coding RNAs
Lys	Lysine
MACS2	Model-based Analysis of ChIPSeq
Me	Methylation
mESC	mouse Embryonic Stem Cells
MLL	Mixed-Lineage Leukaemia, mammalian homolog of TRX

mock IP	Immunoprecipitated in the absence of antibody
mRNAs	messenger RNA
mRNA-seq	messenger RNA high-throughput sequencing
Nipbl	Nipped-B-like protein, Cohesin Loading factor
nt	Nucleotides
Oct4	Octamer-binding transcription factor 4
PCR	Polymerase Chain Reaction
PIC	Pre-Initiation Complex
polyA	poly Adenylation
PRC	Polycomb repressive complex
PROMPTs	PROMoter uPstream Transcripts
P-TEFb	positive transcription elongation factor
Rbp	Retinol binding protein
RefSeq	Reference Sequence
RNA	Ribonucleic acid
RNAi	RNA interference
RNAPII-CTD	RNA Polymerase II C-Terminal Domain
RNAPIIS2p	RNA Polymerase II S2 phosphorylated
RNAPIIS2u	RNA Polymerase II S2 unpshopshorylated
RNAPIIS5p	RNA Polymerase II S5 phosphorylated
RNAPIIS7p	RNA Polymerase II S7 phosphorylated
RPKMs	Reads Per Kilobase of exon per Million reads mapped
RSEM	RNA-Seq by Expectation-Maximization
SE	Super enhancer
Ser	Serine

SHH	Sonic Hedgehog
STAR	Spliced Transcripts Alignment to a Reference
STARR-seq	Self Transcribing Active Regulatory Regions sequencing
TES	Transcription End Site
TF	Transcription Factor
TFBM	Transcription Factor Binding Motif
TFIIH	Transcription factor II Human
Thr	Threonine
TPMs	Transcripts Per Million
TSS	Transcription Start Site
Tyr	Tyrosine
UCSC	Univeristy of California Santa Cruz
WAP	Whey Acid Protein

1. Introduction

All known biological life depends on genetic information, hence the expression of genomes is tightly regulated. Genes comprise regions of the genome that are transcribed to mRNA and code for proteins, and occupy only a fraction (~3% if only the coding region is considered, 40% from transcription start site (TSS) to transcription termination site (TES)) of the entire human genome (Dunham et al., 2012). The remaining genomic sequences can have structural functions, e.g. telomeres and centromeres, transcribe structural RNAs (rRNAs, tRNAs), or be occupied by regulatory regions which have the function of regulating the expression of specific genes at specific times. Gene regulation is pivotal to assure cell identity, proper development (Johnson et al., 2018; Osterwalder et al., 2018), and reaction to stimuli (De Santa et al., 2010; Oishi et al., 2017) with regulatory regions activating or enhancing gene activity. In addition, transcription at genes and the downstream processes are tightly regulated. As can be appreciated in the biological world, different regulatory systems cooperate to assure the right timing and expression levels for each gene that will lead to the specific phenotype of every single cell.

1.1 Regulatory elements

1.1.1 Enhancers

Regulatory regions are scattered through the genome and were described to exert different functions. Enhancers are one such regulatory region which can enhance or activate transcription of specific target genes (Heinz et al., 2015). Enhancers were found in bacteria and eukaryotic cells and are extensively studied as major players in gene-expression regulation. Their mechanisms of action are not fully understood, but the formation of a chromatin loop that brings in proximity enhancers and target genes is one of them (Beagrie and Pombo, 2016). Enhancers act as dock of transcription factors (TFs), co-factors such as Cohesin or Mediator, RNAPII, and are marked by specific histone modifications (HMs) (Calo and Wysocka, 2013) (Fig 1.1).

Enhancers can reside very far from their target genes; a historical example is the Sonic

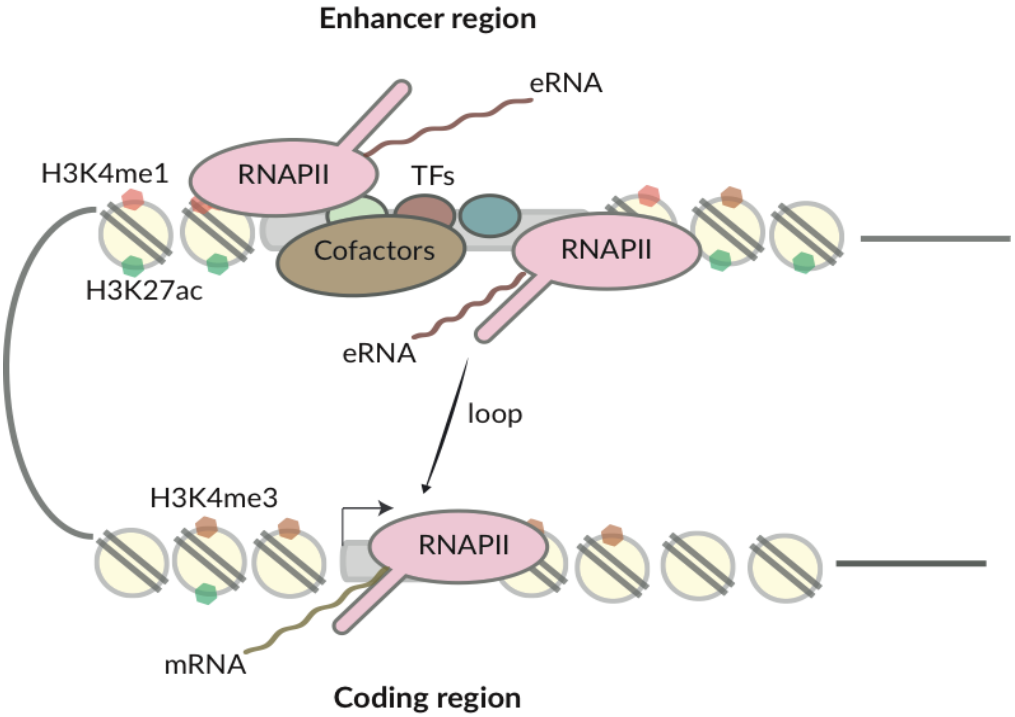


Fig 1.1: Schematic of enhancer regions features

Hedgehog (SHH) locus, spanning ~900kb, where enhancers regulating the SHH gene are located inside a SHH intron (for neuronal expression) and as far as 849kb, inside a neighbouring gene (for limb expression) (Anderson et al., 2014). It has been calculated that every cell can have ~10000 – 50000 enhancers (Andersson et al., 2014a; Arner et al., 2015; Heintzman et al., 2009b; Nord et al., 2013), though even more interesting is that enhancers will not be active at the same time in the same cells (Andrey et al., 2017). As an example, Sonic Hedgehog gene is expressed during development in different cell types, such as lungs or limbs. The specificity of its expression is achieved by the activity of enhancers, which is restricted for cell types. A SHH enhancer active in limbs will be silent in lungs and vice-versa (Anderson and Hill, 2014) (Fig 1.2).

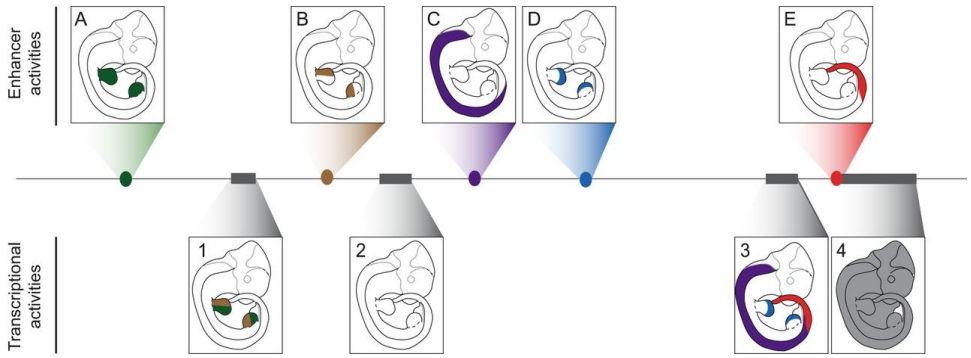


Fig 1.2: Schematic of enhancer and gene activity. A hypothetical locus containing five enhancers (coloured ovals; A-E) and four genes (grey boxes; 1-4) is shown. The regulatory activities of enhancers A-E in embryonic day 10.5 mouse embryos are represented by coloured shading in the diagrams above, and the transcriptional activities of genes 1 to 4 are shown in the diagrams below. A comparison of enhancer activities and gene expression, marked with the same colour code, shows that enhancers A and B contribute to gene 1 transcription, and that enhancers C-E contribute to gene 3 transcription. Gene 2 is repressed in all tissues at this embryological stage, whereas gene 4 displays a ubiquitous expression pattern. From (Andrey and Mundlos, 2017).

Importantly, enhancer deletions or ectopic activity(Lupiáñez et al., 2015), and enhancer mutations (Lettice et al., 2018) are associated with diseases, such as a subset of b-talassemias (Van der Ploeg, 1980) and developmental diseases (Johnson et al., 2018).

1.1.2 Transcription factor binding at enhancers

TFs are found at enhancers and are important for enhancer identity and regulation. Transcription factors are proteins which bind to a specific sequence of DNA, the binding motif (TFBM), and recruit co-factors that can: a) modify the chromatin environment (remodeller of chromatin) and b) recruit other factor such as Cohesin or Mediator that are implicated in chromatin loop formation (Kagey et al., 2010). The influence of transcription factors on genomic regulators is well studied in different cells types and upon various stimuli (Heinz et al., 2010; Zaret and Carroll, 2011). Transcription factors are known to bind promoters of coding genes and regulatory regions, however, some TFs bind preferentially to regulatory regions over promoters, such as FoxA1 and Pu.1 in activated macrophages (Gosselin et al., 2014).

Some TFs are able to bind closed chromatin (Zaret and Carroll, 2011), these “pioneer” transcription factors bind compact chromatin and recruit factors that help opening the locus, facilitating other TFs binding and enhancer activation. Pioneer transcription factors can be important to define the identity of the cell, as well as a response to stimuli, such as Pu.1 (Ghisletti et al., 2010). After pioneer TFs binding, regulatory regions become nucleosome depleted and accessible to factors involved in chromatin remodelling. In this way, enhancers acquire marks specific of their state of activation.

1.1.3 Enhancer activation states

Enhancers can be classified in different states of activation based on combinations of histone modifications, Transcription Factors (TFs) binding, and DNA modifications (Choi et al., 2014).

Enhancers can be classified into different states: inactive, active, primed, poised (Calo and Wysocka, 2013), super (Hnisz et al., 2013; Whyte et al., 2013) and non-canonical (Pradeepa et al., 2016) enhancers . While the exact role of different enhancer states and their definition is currently investigated and debated, there is some consensus on different chromatin characteristics of each activation state. Moreover, for some classes, such as primed and active, there is not direct switch, but more likely a gradual transition from one class to the other.

Historically, enhancers are defined as regions of DNA outside of promoters enriched for H3K4me1 and depleted of nucleosome (Calo and Wysocka, 2013). It was recently found that H3K4me1 marks not only active enhancers, but also primed, poised (Calo and Wysocka, 2013), super (Hnisz et al., 2013) and non-canonical (Pradeepa et al., 2016) enhancers, and only being depleted only from inactive enhancers. H3K4me1 therefore is a general mark of putative regulatory regions, and not sufficient to classify their state precisely.

Active, Primed and Inactive enhancers

Active enhancers are the ones engaged in enhancing target gene expression. Active enhancers are marked by H3K4me1, H3K27ac and show high abundance of TFs binding (Calo and Wysocka, 2013). It is currently understood that the binding of transcription factors is of high importance for enhancer activity (Spitz and Furlong, 2012). Active enhancers were shown to come in proximity to their target genes via chromatin looping and to modulate target gene expression (Beagrie and Pombo, 2016; Krivega and Dean, 2012). P300, the acetylase that acts on H3K27 among other targets, is also present at active enhancers, however it can also be found at less active ones (Creyghton et al., 2010).

Primed enhancers are regions of the genome which are ready to be activated, however are not exerting a specific function yet. Primed enhancers are depleted of nucleosomes and marked by H3K4me1(Calo and Wysocka, 2013). The priming of enhancers was observed in ESC and in differentiated cells (Liber et al., 2010; Xu et al., 2009).

Inactive enhancers are unmarked regions of the genome, which act as enhancers in another cell type.

Poised enhancers

Poised enhancers repress their target genes, are mainly described in ESCs and are marked by H3K27me3 (Creyghton et al., 2010; Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011; Zentner et al., 2011), a histone modification deposited by Polycomb Repressive Complex 2 (PRC2). Poised enhancers have been described in mouse, humans, and drosophila (Creyghton et al., 2010; Koenecke et al., 2017; Rada-Iglesias et al., 2011), are repressed in ESCs and active when a differentiation path is taken by the cell. Neuronal target genes of poised enhancers are poorly transcribed in mESC, however, when cells start to differentiate to neuronal lineages, poised enhancers lose H3K27me3, acquire H3K27ac, and their target genes get activated (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). Poised enhancers tend to be in closer proximity to transcription start sites (TSSs) compared to active and primed enhancers, are more enriched in GC, and are more evolutionary conserved than other classes of enhancers(Aran et al., 2016; Rada-Iglesias et al., 2011).

Poised enhancers were also defined in mESCs as regions of the genome enriched in 5-formylcytosin (5fC) or 5-hydroxymethylcytosine (5hmC)(Choi et al., 2014), and transcriptionally silent. 5hmC is also present at bivalent promoters, containing both H3K4me3 and H3K27me3 marks (Matarese et al., 2011), which suggests a similarity between repressed genes and repressed enhancers.

Recently, a study in Drosophila challenged the view of poised enhancers as Polycomb targets, arguing that the Polycomb signal originated from nearby genes. Therefore, acquiring the H3K27me3 mark would be a side-effect of the proximity effect of H3K27me3 broad deposition(Koenecke et al., 2017). The biological relevance of this observation is yet to be understood.

Strikingly, it was shown for a number of poised enhancers that they contact their target genes already in mESC, where they exert a repressive function on target genes. Fascinatingly, Cruz Molina *et al.* showed that target genes remain silenced in mESCs and are activated during differentiation, when poised enhancers keep the contact and switch to an active state. Interestingly, if contacts are disrupted in mESCs proper gene transcription upon differentiation is lost(Cruz-Molina et al., 2017).

Super enhancer/ Stretched enhancers/LCR

In 1980 a region in the β-globin locus was described to be DNAseI hypersensitive and, if deleted, caused a special form of thalassemia(Van der Ploeg, 1980). This region was called

Locus Controlled Region (LCR) and it is now understood as a region dense in enhancers(Li et al., 2002). More recently, clusters of enhancers were described by different groups and in various organism(Parker et al., 2013; Whyte et al., 2013), re-named Super Enhancers (SE)(Whyte et al., 2013), and patented. SEs are regions of the genome that greatly enhancer target gene expression, bound by TFs, and highly enriched in H3K27ac, Med1 and Brd4 binding. Among their targets are cancer genes(Hnisz et al., 2015; Hnisz et al., 2016), cell identity genes, and developmental genes(Whyte et al., 2013). SE were shown to be involved in disease and cancer progression(Ko et al., 2017), making them biologically relevant.

SE in the Whey Acid Protein (WAP) locus were described as a cohort of enhancers acting synergistically acting in a hierarchical fashion (Shin et al., 2016). However, deletion of a portion of SE in the beta-globin locus demonstrated that enhancers inside these regions can act independently(Hay et al., 2016) .

Non-canonical enhancer

Recently, two new histone modifications, H3K64ac and H3K122ac were characterised at regulatory regions and described as marks of a new and previously unknown class of enhancers, called non-canonical enhancers(Pradeepa et al., 2016). H3K122ac was first described to be present at TSSs, to co-occur with active chromatin marks such as H3K27ac, H3K4me1/3, but not with H3K36me3 (marking active coding regions) or H3K9me3 (marking constitutive heterochromatin). Failure in H3K122 acetylation impairs rapid transcriptional activation in yeast. H3K122 acetylation is catalysed by CBP and P300, the latter being an acetylase responsible also for H3K27 acetylation at enhancers. H3K122ac is found at distal regions with a potential enhancer activity; higher levels of H3K122ac after estrogen stimulation correlate with increased eRNA transcription(Tropberger et al., 2013).

Interestingly, H3K122ac doesn't correlate completely with H3K27ac distribution at distal regions(Tropberger et al., 2013). In fact, Pradeepa and colleagues(Pradeepa et al., 2016) showed that H3K122ac, often together with H3K64ac, can occur at regulatory region depleted of H3K27ac, but marked by H3K4me1. The authors called these regions "non-canonical" enhancers. Non-canonical enhancers were shown to be associated with genes important in stem cell maintenance and brain morphogenesis, to transcribe high levels of exosome-sensitive eRNA, and to be bound by high levels of P300. Interestingly, a subset of these non-canonical enhancers was enriched in H3K27me3 and associated with genes involved in development and differentiation. In the same work it was observed that non-canonical enhancers are more

enriched in protein such as CTCF and RAD21, which may point to a regulatory role involving chromatin loops and boundaries(Pradeepa, 2016).

1.1.4 Transcriptionally active enhancers

Enhancers were found to transcribe a newly identified class of RNAs, eRNAs (enhancer RNA)(Kim et al., 2010; Kim et al., 2015; Kaikkonen et al., 2013). Transcriptional active enhancers show mainly a bidirectional transcription(Andersson et al., 2014a) of short-lived RNAs (Andersson et al., 2014a; Pradeepa et al., 2016) and were classified as a complementary state to active enhancer state. eRNAs are usually not detected through polyA-RNAseq, probably due to a low level of poly-adenylation (estimated as ~10% by Andersson *et al.* 2014 (Andersson et al., 2014a)) and low abundance. Some studies report eRNAs to be unspliced and not polyadenylated (Andersson et al., 2014a), others find eRNAs to be spliced and with a polyA tail (De Santa et al., 2010), making difficult to draw a general rule. They are therefore detected with techniques that do not rely on poly-T enrichment, such as GRO-seq, total RNA (De Santa et al., 2010; Koch et al., 2011) and CAGE (Andersson et al., 2014a; Arner et al., 2015).

Reports show that transcription at enhancer starts from a central point marked by a P300 peak. Within this region two transcription initiation events may take place, which were found to be separated by 180bp (Andersson et al., 2014a). This observation resemble the divergent transcription described at promoters of active genes (Duttke et al., 2015). Upon exosome knockdown, which is responsible for the degradation of not mature RNA, the relative frequency of eRNA detection increases (Andersson et al., 2014a), suggesting that eRNAs are fast degraded upon transcription and that the Exosome plays an important role in transcription regulation at enhancers. Interestingly, disruption of chromatin loop via Cohesin knock down was shown to not affect eRNA production (REF Ing-Simmons 2015 Genome Research), showing that eRNA transcription is not dependent on chromatin organisation.

eRNA transcription is now a common approach to identify active enhancers (Andersson et al., 2014a; Henriques et al., 2018). SE were also found to be robustly transcribed (Blinka et al., 2016; Hnisz et al., 2015), possibly due to their active state coupled with high transcription factors occupancy and RNAPII binding (Blinka et al., 2016).

1.1.5 Techniques to identify enhancers

The importance of enhancers as regulatory elements makes their identification in the cell of interest an important field, continuously evolving. The approaches differ and vary both in methodology and in quality. On the one side, enhancers are identified using bioinformatics

approaches, usually feeding a number of datasets into a software that computes the likelihood of a specific region in the genome to be a regulatory element (Bu et al., 2017; Chen et al., 2012; Ernst and Kellis, 2012). On the other side, enhancers can be detected with the use of specific datasets in a more digital way: if specific marks are present, the region will be categorized as an enhancer.

ChIP-seq experiments are usually performed to understand the binding pattern of TF and to identify putative enhancer regions. The drawback of this approach is the limited knowledge we have about specific TFs for cell types, especially rare and not well characterise ones. In addition, the specificity of transcription factors binding to identify enhancer responsive to a stimulus or a differentiation pathway needs prior knowledge of the specific factor involved.

Transcription factors binding motifs (TFBM) recognised by families of TFs can be used to identify putative regulatory regions in various cell types. However, TFBMs can be masked either by a specific structure or by other marks and therefore not recognised by their TF in all cell types, which adds another level of gene regulation. The sole analysis of binding motifs across the genome, even if associated with a TF expression analysis, is therefore not sufficient in most cases to understand if that specific regions would be bound and active in the cell type of interest (Slattery et al., 2014).

To identify SE, enhancers are firstly identified with canonical marks and then ordered by the enrichment for one of the specific features of SE (H3K27ac, Med1): the enhancers, which are highly enriched for the feature would be categorized as SE (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). This type of approach strongly relies on the clustering methods of normal enhancer prior to the enrichment analysis.

Enhancers have also been identified based on bidirectional transcription of eRNAs outside promoters with techniques such as GRO-seq (Core et al., 2014; Danko et al., 2015) and CAGE (Andersson et al., 2014a; Arner et al., 2015). Enhancers identified with these approaches are defined as genomic regions outside annotated promoters from which RNAs originate. The techniques used to detect RNAs make some assumptions on the nature of the eRNAs: CAGE technique, for example, detects only capped enhancers; GRO-seq has an *in vitro* step which can influence transcription at enhancers, which is less stable than the one at genes (Henriques et al., 2018), and cannot distinguish between elongating and non-elongating RNAPII. These approaches aim to identify transcriptionally active enhancers, however their sensitivity is still debated and it is not clear whether they can identify only active enhancers, or also enhancers in other activation states.

The comparison of the number of detected enhancers in similar cell types with different approaches clearly shows how approaches based on different assumptions lead to the detection of a different number of regions (REF Fig 1.3). The variety of techniques, coupled with our limited knowledge of enhancer activation states, make the identification of a complete set of enhancers in a cell challenging.

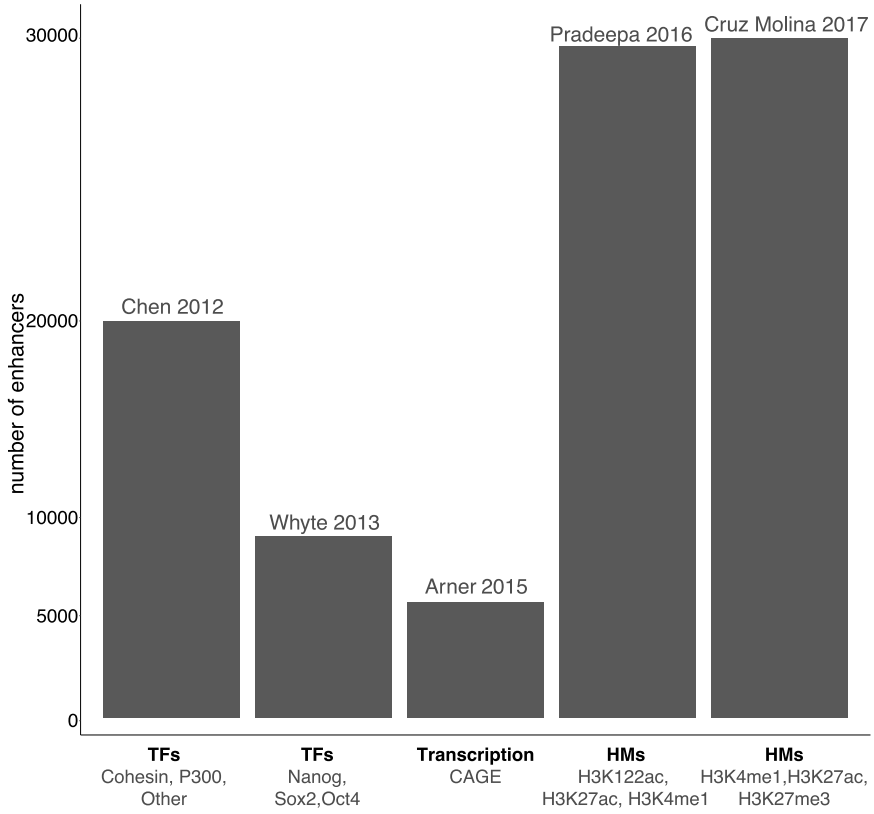


Figure 1.3: Number of enhancer regions identified with different approaches. Approached used in the identification on the x-axis, publication on top of the bar. Number shown in the graph are as indicated in the original publications.

1.1.6 Mechanism of enhancer activity

Enhancers have been shown enact transcriptional change via different mechanisms. One of the most known is the looping model, where enhancers are linearly distant from their target and are brought in proximity via a loop of the chromatin (Krivega and Dean, 2012). The looping mechanism is thought to be, at least in part, dependent on Mediator and Cohesin proteins (Kagey et al., 2010). Interestingly, enhancers can be in contact with their target genes prior to activation(Cruz-Molina et al., 2017), or with genes on other chromosomes (Markenscoff-Papadimitriou et al., 2014). A study on the SHH locus showed with 3C technologies and DNA-RNA FISH that contacts between enhancer and promoters are specific for the cell type and are in place before the target gene expression. While contacting regions can be also primed or repressed, expression of target gene is achieved only when contact is established (Amano et al., 2009).

Despite the chromatin looping being a well-established phenomena, the functional role of the contact or proximity of enhancers and target genes it is not well understood. It is possible that enhancers create an environment rich in TFs, co-factors and RNAPII close to the gene, resulting in enhanced transcription frequency or robustness. Moreover, enhancers can act on the transcriptional burst, making the transcription event more stable and on the release of RNAPII from promoter proximal pausing (Chen et al., 2017; Liu et al., 2013; Schaukowitch et al., 2014b).

Different classes of enhancers in different tissues were proven to be in contact with their target genes (de Laat and Grosveld, 2003; Li et al., 2013; Sanyal et al., 2012), also prior to activation (Cruz-Molina et al., 2017). Moreover, it was shown that enhancer activity is confined in chromatin domains that, if compromised, will lead to ectopic gene expression (Lupiáñez et al., 2015). 3C technologies, such as 4C, HiC, ChIAPet, have been recently used to understand enhancer-promoter contacts. This led to the interesting observation that more than one enhancer is in contact with the same target gene (Markenscoff-Papadimitriou et al., 2014). Recently our lab developed a new ligation-free technique that enables to study multiple interaction genome-wide, Genome Architecture Mapping (GAM). GAM made possible to show that SE are highly enriched in contacts with multiple other super enhancers and active genes in mESC(Beagrie et al., 2017).

Another proposed model of enhancer mechanism, the “tracking” model, describes enhancers as docks for RNAPII that can “track” chromatin in either direction until they encounter a promoter to start the productive transcription (Bulger and Groudine, 1999; Vernimmen and Bickmore, 2015). In the tracking model, RNAPII binds at a regulatory region upstream the promoter, tracks the chromatin without producing mature transcripts and, once it reaches the promoter, starts an efficient mRNA transcription. It is not clear if the enhancers using the tracking model mechanism have different genomic or epigenetic features compared to the ones looping to the target, or if RNAPII is present in a unique forms at these enhancers.

Another mechanism through which enhancers could act indirectly on target genes is via their transcripts. eRNAs were shown to be transcribed before their target gene(Arner et al., 2015; Li et al., 2015) or to correlate with target gene expression(Kim et al., 2010; Lam et al., 2013; Kaikkonen et al., 2013). For example, it was shown in immune cells that upon stimulus, enhancers are bound by the TF PU.1, start transcription and sub sequentially the target gene is transcribed (Kim et al., 2010). A diminished concordant expression of eRNA and mRNA was observed in mouse brain with Huntington disease, compared to healthy brain (Le Gras et al.,

2017), and RNAi against eRNAs could diminish target gene transcription (Lai et al., 2013; Lam et al., 2013; Li et al., 2013; Melo et al., 2013), suggesting a role of coordinate transcription for normal target gene expression. In contrast to these results, a report showed with single molecule FISH that eRNA expression is not necessary for mRNA expression. eRNA and mRNA in this system are rarely co-expressed, and eRNA expressions happens in the allele not involved in the chromatin loop (Rahman et al., 2016). Another study showed that insertion of a premature termination cassette, which impaired eRNA expression but not transcription at enhancers, didn't have any effect on target gene expression (Engreitz et al., 2016).

Depletion of eRNAs does not seem to impact the chromatin environment of enhancers, as levels of H3K27ac, H3K4me1 and TFs binding remain unchanged after RNAi, suggesting that eRNAs do not have a role in enhancer chromatin environment maintenance (Blinka et al., 2016). Efforts in recent years show that some eRNAs can have roles in stabilizing the loop (Lai et al., 2013; Li et al., 2013), however transcription of eRNA was suggested to be not crucial for loop formation, which, once established, persists upon transcription inhibition (Hah et al., 2013). The exact roles of transcription at enhancers and eRNAs are still to be understood.

1.1.7 Methods to study enhancer activity and find enhancer target genes

Historically, an enhancer is a region of the genome that, inserted in a reporter plasmid, will generate a reporter signal, due to the action of the inserted tested enhancer in a vector with a reporter gene under a weak promoter, and is an approach that is widely used to test if a region acts as an enhancer. For example, in the VISTA enhancers' atlas (Visel et al., 2007) different putative regulatory regions are tested for their expression patterns in embryos and organs in human and mouse with a reporter assay. High-through reporter assays techniques were recently developed, greatly advancing progress. For example, the STARR-seq (Self Transcribing Active Regulatory Regions sequencing) method tests loci for enhancer activity shearing the genome and using RNA-seq to screen for enhancer activity (Arnold et al., 2013).

Although reporter approaches have proven extremely valuable to understand which regions would function as enhancers, as well as to identify and study the minimal region (the minimal number of base pairs) that drives reporter expression (Milewski et al., 2004; Small et al., 1992), reporter assays do not take into account the effect of the 3D organisation of the chromatin. This is especially important as plasmids are artificial "environments" that do not resemble the original chromatin region.

Taking advantage of chromatin-contact information, such as 3C technology derived, could help identify regions of the genome in contact with a specific gene of interest. Contact-mapping methods proved useful to understand the proximity of different regions of the genome, however Loss of Function experiments are often needed to prove the function of regulatory regions. Interesting experiments using CRISPR-Cas9 methodology showed different effects of enhancer KO. in Blinka et al. (Blinka et al., 2016) the authors knocked-out 3 SE acting on Nanog, and observed diverse outcomes. One deletion had effect on both Nanog and Dppa3, a neighbouring gene; the second deletion did not have any effect in the locus; the third was recovered only in monoallelic clones with a 50% reduction in Nanog expression. CRIPR-Cas9 mediated KO was also used by Cruz Molina and colleagues to show that deletion of poised enhancers leads to a failure in activation of the target genes upon differentiation.

A parallel approach to CRISPR-Cas9 to sample unbiasedly a region and determine target genes of regulatory regions is via the use of retro-transposon hopping (Anderson et al., 2014). The expression of genes of interest is checked while the retro- transposon occupies a diverse position: when a variation in gene expression is detected, the region occupied by the transposon and the genes of interest are matched.

eRNAs transcription correlation with target gene expression was also used to infer target genes during time courses (Arner et al., 2015). This approach can estimate the target genes, however can only be applied to time courses with numerous time points.

How enhancers are able to recognise their target genes in the 3D space of the nucleus, and why some genes are uninfluenced by close enhancer activity is still under debate.

1.2 Transcription

Coding DNA sequences are transcribed to RNA by the enzyme RNA Polymerase II (RNAPII). RNAPII is recruited to gene promoters to form the Pre-Initiation Complex (PIC) together with other proteins such as Mediator (Hahn and Young, 2011; Thomas and Chiang, 2006), and start transcription. During transcription different enzymes are responsible of the regulation of the process speed and frequency of RNA processing, and of chromatin remodelling another layer of transcriptional regulation. The recruitment of the majority these proteins and their complex interaction is integrated through the C-terminal domain (CTD) of RNAPII and its post-translational modifications.

RNAPII is a well-conserved and highly regulated enzyme and it is composed by 12 (Rbp1-12) subunits in mammals. The main subunit Rbp1 contains a globular domain and an unstructured

domain at its C-terminal domain. The CTD of RNAPII is composed of 52 tandem repetitions in mammals (44 repetition in Drosophila and 26-27 in Yeast) of the consensus heptapeptide Tyrosine1-Serine2-Proline3-Threonine4-Serine5-Proline6-Serine7 (Y1-S2-P3-T4-S5-P6-S7) (Chapman et al., 2008; Yang and Stiller, 2014), which can be extensively post-translationally modified. The post-translational modification of RNAPII-CTD determines its activation state and can act as a recruiter of chromatin remodellers and RNAPII maturation machinery.

1.2.1 RNA Polymerase II states of activation: Serine phosphorylation

Transcription at coding genes is a well-studied process characterised by different stages that comprise initiation and elongation. Every stage of gene transcription is associated with a specific form of RNAPII, which is determine by its post-translational modifications. The most known modifications of the RNAPII CTD are the sequential phosphorylation of the Serine residues 5 (Ser5p), 7 (Ser7p), and 2 (Ser2p) (Akhtar and Gasser, 2007; Tietjen et al., 2010). Ser5p occurs first at gene promoters and recruits the capping machinery (Ghosh et al., 2011), Ser7 is subsequently phosphorylated, an event that is necessary for Ser2 phosphorylation, productive elongation, and recruitment of RNA maturation machinery (Fig 1.4) (Corden, 2013; Gu et al., 2013; Lunde et al., 2010). The specificity of phosphorylation is reflected in their pattern along the gene body, with Ser5p peaking at the promoter and gradually descending, while Ser7p and most notably Ser2p are more enriched at the gene body and peak after the termination end side (TES). Among the specific enzymes responsible for the phosphorylation of the CTD residues, the cyclin-dependent kinase (Cdk) family is well characterized (Fisher, 2017) and responsible for the transitions between different stages of transcription.

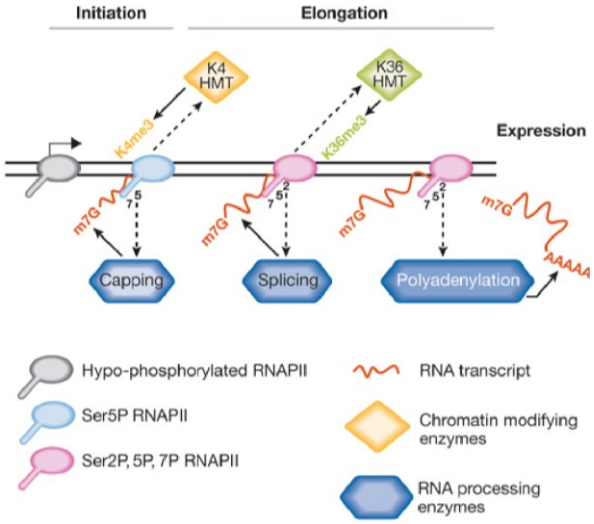


Fig 1.4: Integration of carboxy- terminal domain modifications with chromatin structure, RNA processing and Polycomb repression. Distinct modifications of the CTD assist in recruiting different chromatin modifying enzymes and RNA processing factors. At active gene promoters, the phosphorylation of RNAPII at Ser 5 (Ser 5P) recruits HMTs to methylate H3K4 and the RNA capping machinery to add an m7G cap to nascent RNAs. Ser 2P creates an elongating polymerase that recruits the HMTs responsible for trimethylation of H3K36 and RNA processing factors. The mRNA, which is released after termination, is stabilized by its cap and poly(A) tail, thereby promoting mRNA transport and protein expression. From (Brookes and Pombo, 2009)

Serine phosphorylation is recognized by protein complexes which are implicated in RNA maturation and chromatin remodelling (Bentley, 2014; Buratowski, 2009; Egloff and Murphy, 2008; Hsin and Manley, 2012). For example, Ser5p recruits the RNA capping machinery (Fabrega et al., 2003), while Ser2 phosphorylation recruits the polyA and the splicing machinery (David et al., 2011; Proudfoot et al., 2002).

1.2.2 Other RNA Polymerase II modifications

Alongside phosphorylation of the Serine residues, RNAPII-CTD can be modified in other residues. For example, Thr4 residue of the CTD hepta-peptide repeats can also be phosphorylated, and this modification is associated with elongation at active genes(Hintermair et al., 2012) . Tyr1 can also be phosphorylated and its profile resembles the one of Ser2p: gradually increasing at the gene body, however it drops before the transcription termination site (TES), and inhibits the recruitment of the termination machinery (Mayer et al., 2012b). Immuno-precipitation experiments showed the concomitant phosphorylation of Ser5 and Ser7, together with Tyr1, but not of Ser2, on the same CTD (Mayer et al., 2012b). Ser2p and Tyr1p could therefore be mutually exclusive. Tyr1p was reported at bidirectional promoters and at enhancers (Descostes et al., 2014; Mayer et al., 2012b).

The distal part of the CTD can present modification of the consensus motif, which are more common in vertebrates (Dias et al., 2015). Non-consensus Lys7 can be acetylated and this modification is associated with active transcription and elongation (Schröder et al., 2013). Recently, our lab and others showed that Lys7 can also be mono/di/tri methylated and that these modifications are conserved from unicellular eukaryotes to mouse (Dias et al., 2015). Lys7me is linked with early stages of transcription and can co-occur with Ser5p and Ser7p, the balance between Lys7 methylation and acetylation levels were shown to be predictive of the transcriptional state of genes. Lys7mes is remarkably absent at Polycomb-repressed genes, suggesting a specific association with active genes and mRNA production (Dias et al., 2015).

1.2.3 Transcription Regulators

Transcription at genes is a tightly regulated process that involves different players and steps. First, RNAPII binds at the promoter, and is phosphorylated on its Ser5 residue by Cdk7 from the TFIIH complex. 20-60nt downstream the TSS RNAPII pauses (Adelman and Lis, 2012). This pausing is mediated by Nelfa and DSFI/Spt5 that are dismissed by the P-TEFb complex. The recruitment of the P-TEFb complex at genes is mediated by Brd4 (Jang et al., 2005): Brd4 was shown to be pivotal for the formation of the elongation complex (Winter et al., 2017). P-TEFb complex phosphorylates RNAPII and releases it from the promoter proximal pausing. P-TEFb

complex contains the kinase Cdk9, which phosphorylates Nelfa and DSFI/Spt5 (Adelman and Lis, 2012). After phosphorylation Nelfa detaches from RNAPII, while DSFI/Spt5 remain associated with RNAPII during elongation. Cdk9 also phosphorylates Ser7 and Ser2 (Baumli et al., 2008) and it is required for Thr4 phosphorylation (Hsin et al., 2011), bringing the transcription to its elongating phase, at the end of the process transcription can be reinitiated. This step is controlled by the Med12-Cdk8 complex, which negatively affects reinitiation. Cdk8 was suggested to inhibit the Med13-mediated RNAPII recruitment and therefore to negative influence the re-establishment of another round of transcription (Knuesel et al., 2009). The Mediator complex is composed by up to 26 subunits in mammals and act as an integrator of TFs to present them to RNAPII (Allen and Taatjes, 2015; Conaway and Conaway, 2011). Mediator has shown to have wide roles, being implicated in gene regulation, including initiation and elongation, and enhancer-promoter contact via chromatin looping (Allen and Taatjes, 2015).

1.2.4 Polycomb and poised genes

Some genes were described to be in a special chromatin state, called a bivalent state. Bivalent genes show marks of active and repressed chromatin, H3K4me3 and H3K27me3 respectively (Azuara et al., 2006; Bernstein et al., 2006b).

Our lab showed that RNAPII phosphorylated in its Ser5 residue is present at bivalent genes together with the Polycomb Repressive Complex (PRC) 1 and 2 (Brookes et al., 2012; Stock et al., 2007). Polycomb genes bound by RNAPII are associated with rapid activation upon stimuli (Voigt et al., 2013). Recently, colleagues in the lab showed that RNAPII/Polycomb genes (poised genes) are also found during differentiation and are very interestingly associated with genes encoding for TFs important for trans-differentiation (Ferrai et al., 2017). Importantly RNAPII and Polycomb were shown to be present at the same allele, as consequential-ChIP experiments demonstrated (Brookes et al., 2012; Ferrai et al., 2017). Some of the Polycomb repressed genes were found with Ser5p together with activating Ser7p and also producing low abundant RNA (Ferrai et al., 2017), however their functions are currently unknown. Interestingly, Ser5 phosphorylation at Polycomb genes is mediated by the Erk2 kinase (Tee et al., 2014). Poised genes are also regulated by Utf1 that has the double function of preventing H3K27me3 spreading and of regulating the levels of unnecessary RNAs produced at poised genes (Jia et al., 2012). Moreover, RNAPII-CTD at poised genes is in an unknown conformation distinct from active genes: the 8WG16ab, which recognises the un-phosphorylated form of Ser2, together with Ser5p at active genes, is unable to bind poised genes, suggesting a specific conformation of the CTD at poised genes (Brookes and Pombo, 2012).

1.2.5 RNAPII at enhancers

RNAPII is present at enhancer and transcribes eRNAs, the amount of which correlates with enhancer activity (Kaikkonen et al., 2013).

Some studies analysed different forms of RNAPII at enhancers with the use of antibodies specific for RNAPII modifications. For example, Ser5p was found at enhancers in mESC (Cruz-Molina et al., 2017) and in differentiated cells (De Santa et al., 2010). An interesting publication showed that during human endodermal differentiation RNAPII phosphorylated in Ser5, but absent or lowly in Ser7 co-localises with enhancers bound by LIF-1 (Estarás et al., 2015). 3D contact studies via ChIA-PET on RNAPII (Reeder et al., 2015) showed the presence of the enzyme both at genes and at regulatory regions, contacting each other. This study also showed, via network analysis of contacting regions in the genome (Pancaldi et al., 2016), that regions enriched in contacts and outside promoters have RNAPII enrichment, with remarkably high levels of Ser2p.

An ongoing dispute is whether Ser2 is phosphorylated at enhancers. Some reports show its presence (Pancaldi et al., 2016), while other detect little to no enrichment (Koch et al., 2011). This discrepancy could be due to different reasons, firstly different antibodies can recognise different epitopes on the CTD. Secondly, the regions under examination in the different works may not share the same features and functions, or may even be in different activation states.

RNAPII at SE correlates with Med1, Brd4 levels (Lovén et al., 2013), and SE activity. RNAPII was also found at poised enhancers when they transition to the active state (Rada-Iglesias et al., 2011; Zentner et al., 2011). Moreover, RNAPII at enhancers is sensitive to a Cdk9 inhibitor (Flavopiridol) (Hah et al., 2013), which suggests that RNAPII transcription at enhancers is regulated at least in part via proteins known to regulate transcription at genes. Recently, Henriques and colleagues showed that transcription at enhancers in Drosophila is generally unstable compared to the one at genes and regulated by Spt5. Hence, the authors conclude that the elongation machinery is fundamental for transcription at enhancers (Henriques et al., 2018), however it remains to be determined if this finding holds true in mammalian cells.

1.2.6 Differences and similarities between enhancers and promoters

Some reports show that promoters and enhancers have a good degree of similarities: such as RNAPII and TF binding. Active promoters were believed to be enriched in H3K4me3 over H3K4me1 at enhancers (Heintzman et al., 2009a), however recently, this view was challenged by a study on transcriptionally active enhancers in Drosophila, that show active transcription

correlates with higher levels of H3K4me3 at enhancers (Henriques et al., 2018) Furthermore, the observation that active promoters can be bi-directionally transcribed (Duttke et al., 2015) and that bi-directionality of promoters correlate with their evolutionary age and specialisation (Jin et al., 2017), suggest that enhancers and promoters share more features than previously imagined. Tyr1p is also found at antisense promoters and actively transcribed enhancers, in human cells however, it is depleted at coding gene promoters (Descostes et al., 2014; Hsin et al., 2014).

Promoters can act as enhancers of other genes in reporter assays and contact them via formation of chromatin loops (Dao et al., 2017), and that some enhancers can function as weak promoters (Mikhaylichenko et al., 2018). In particular, eRNAs share similar features with PROMoter uPstream Transcripts (PROMPTs). PROMTs are unstable transcripts originating from bidirectional promoters, which showed exosome and Flavopiridol (Flynn et al., 2011) sensitivity and are mainly not-polyadenylated (Andersson et al., 2014a). PROMPT regions are bound by RNAPII in different forms of activation, comprising also the fully elongating form RNAPIIS2p (Preker et al., 2011). It has been suggested that early termination sites will cause premature termination and RNA degradation, a mechanism that was suggested for eRNAs as well (Grzechnik et al., 2014). Moreover, it has even been shown that promoter upstream regions can act as enhancers (Dao et al., 2017; Serfling et al., 1985; Zabidi et al., 2014).

1.2.7 Studying RNAPII

The majority of studies dealing with RNAPII are conducted using specific antibodies that recognise the modified residues of the CTD, followed by pull-down and next generation sequencing (ChIP-seq). While this approach has proven very valuable, some aspects should be taken into account when analysing these data.

An important aspect is the choice of the antibody itself. A number of commercially available antibodies targeting different modifications of RNAPII can be found, however not all of them would show the same profile at genes for the same modification. For example, among two different antibodies targeting Ser5p, 4H8 and E8 (Dias J, personal communication) only one shows RNAPIIS5p presence at gene bodies (4H8). The reason is the specificity of epitopes recognition of the different antibodies, specific epitopes can be marked by other close modification or by tertiary protein folding of the CTD.

Interestingly, this extreme specificity of antibodies made possible to observe a yet not completely understood conformation of RNAPII-CTD at Polycomb genes. RNAPII, when present together with Polycomb at poised promoters, is recognised by antibodies binding Ser5p. However, these regions are not bound by 8WG16, an antibody that recognises the

unphosphorylated form of Ser2. 8WG16ab signal is found at active genes together with Ser5p, which suggests that RNAPII-Ser5p at poised genes have a specific and yet to be understood conformation at the CTD which impairs 8WG16ab binding (Brookes and Pombo, 2012).

In conclusion RNAPII is heavily post-translationally modified and these modifications modulate protein recruitments and, as a consequence, the chromatin state of the loci to which RNAPII is bound and the maturation state of the RNA transcribed.

1.3 Computational approaches

High-throughput methods revolutionised the study of gene regulation, however, vast amounts of data are produced and this brings its own problems in terms of assessing quality.

1.3.1 ChIP-seq Sequencing Quality Check

Chromatin immune precipitation sequencing (ChIP-seq) data recapitulate the occupancy of proteins and histones on genomic DNA. In a typical ChIP-seq experiment, proteins and DNA are cross-linked, the DNA sheared, amplified, and sequenced. At this stage the information is completely digital, coming as a series of nucleotide sequences called reads, which have to undergo a series of steps. Quality checks are pivotal to assure reliability of data. A common issue is that some regions unbound by the protein of interest can be carried on during the purification process to the amplification step (Teytelman et al., 2013).

A first quality check is performed on raw reads coming directly from the sequencer. Reads are tested for quality that can be used to filter out poor quality reads, which are inferred by the sequencer in the amplification clusters; residual adaptor sequences are also removed. Reads are then mapped to a genome of reference, numerous genome mappers are available, which differ in their mapping approach, such as Bowtie2 (Langmead et al., 2009) or Burrows-Wheeler Aligner (BWA (Li and Durbin, 2009)). ChIP-seq reads do not need a splicing-aware algorithm and regions mapping to multiple regions of the genome (multimappers) are usually not considered in downstream analyses. After mapping, ChIP-seq duplicated reads are preferentially removed, because of them being a likely product of PCR duplicates. ChiP-seq reads mapped to the reference genome can be loaded on a genome browser to inspect their quality and for preliminary analysis.

Differential ChIP-seq analysis between datasets can be performed on peaks of enrichment or on raw reads and numerous methods have been developed to perform these types of comparison. However, it has been shown that different approaches have low levels of agreement (Steinhauser

et al., 2016), making it a difficult task to draw conclusions from direct comparison of ChIP-seq data.

1.3.2 RNA sequencing data analysis TPM-FPKM-normalised counts

RNA-seq datasets are derived from the amplification and retro-transcription of species of RNA in the cells of interest. RNA-seq analysis differ from ChIP-seq analysis in some aspects. RNAs can undergo splicing, which should be taken into account by the software which maps RNA-seq reads to the genome of reference. Examples of splicing-aware mappers are STAR (Dobin et al., 2013) and TopHat (Trapnell et al., 2012b). After mapping, duplicated reads are not removed, as these are relative to the amount of the specific RNA species in the cell. Reads originating from total RNA-seq which amplify all the RNA molecules in a cell, including microRNAs, should be analysed for over representation of ribosomal RNA (rRNAs). Quality checks interpreting the percentage of rRNA presence in total RNA datasets serve to understand the quality of the rRNA depletion and the quality of the data.

Reads alignments are used to quantify expression, as it was demonstrated that the abundance of a transcripts correlates with its expression (Trapnell et al., 2012a); expression estimates can be calculated on the isoform level or on the gene level. This is due to the short length of RNA-seq reads, which doesn't permit to distinguish isoforms in all cases. Moreover, RNA-seq analysis can be performed on counted reads or reads estimates. Software using counts on the gene levels are of the such of HTSeq (Anders et al., 2014), while between the software that estimates isoform or gene expression are CuffLinks (Trapnell et al., 2012a) and RSEM (Li and Dewey, 2011). The advantages and disadvantages of each approach are highly debated in the field.

Different expression units exists for RNA-seq are widely used to compare datasets. The most used are Reads or Fragments Per Kilobase of exon per Million reads mapped (R/FPKMs), Transcripts Per Million (TPMs) and normalised counts. FPKMs and TMPs are length-normalised and are scaled for the number of reads sequenced. FPKMs are the number of reads mapping to a region, divided for length all multiplied by 10^9 scaling factor (Mortazavi et al., 2008). TPMs are similar to FPKMs, but are normalised so that the sum of all TPMs in a datasets equals to 1 million. TPM was shown to be proportional to total RNA abundance (Wagner et al., 2012). Normalised counts do not take the length of the gene into account when calculating the abundance and are used by software such as DESeq2 (Love et al., 2014). DESeq2 is used to confront the issues of using different datasets and implements a normalisation step between datasets. DESeq2 calculates the "size factor", which takes into accounts differences in

sequencing depth between datasets, the size factor is used to normalise counts between datasets to permit fair comparison.

1.3.4 Peak finders

Reads containing information on protein and histone occupancy can cluster in domains in the genome, which may be broad or narrow, depending on the nature of the data. The study of these regions of enrichment, called peaks, is well established in the field. Peak finders are algorithms that scan the genome and look for regions of enrichment for the specific mark of interest. Most of the peak finders use a background as reference, as some regions of the genome are likely to be present in a pull-down non-specifically (Teytelman et al., 2013). The use of a background allows the algorithm to detect a specific signal, reflective of the pull-down, over the noise of unspecific regions determined by the background reference. ChIP-seq performed with unspecific antibodies or input chromatin pull-downs are classical backgrounds.

Factors of interest can have different binding-profiles on the DNA. TFs, for instance, tend to have narrow peaks, histone modifications, on the other hand, are broadly distributed. Some peak finders can be informed of the nature of the mark and therefore adjust their search for broad or narrow peaks, such as Bayesian Change-point Model (BCP, (Xing et al., 2012)) and Model-based Analysis of ChIPSeq (MACS (Zhang et al., 2008)). Another layer of complexity arises with mixed peaks, such as RNAPII peakswhich have a narrow peak at the TSS or TES of genes and a broad distribution through the gene body.

Peak finders are widely used and extremely powerful, however raw reads are also a source of information not to be discarded. While peak finders can establish the boundaries of regions of specific enrichments, the amount of enrichment is still defined by the raw reads.

2. Methods

2.1 RNAChIP datasets generation

RNAChIP data in 46C and OS25 mESC clones were generated by KJ Morris and RA Beagrie prior to when I joined the lab and the datasets are currently unpublished. KJ Morris established the RNAChIP protocol and published it in her thesis: "Interplay between Polycomb repression and RNA Polymerase II in Embryonic Stem Cells" (Morris, 2012). OS25 cells are grown in serum+LIF conditions under selection for Oct4 expression, whereas 46C cells are grown in serum free conditions. In brief, cells are sonicated and cross-linked, immune precipitated with 4H8ab against RNAPIIS5p and RNA is extracted. RNA was then sequenced according to illumina's instruction (#1004898 Rev A). Fragment size of 200-450bp were selected.

2.2 Total RNA-seq dataset generation

Total RNA data in 46C mESC clone and 46C mESC differentiated cells in neuronal lineage according to (Ferrai et al., 2017) were generated by AM Fernandes and Carmelo Ferrai. Total RNA was extracted from cells using TRIzol (Invitrogen, Cat# 15596- 018), following manufacturer's instructions. 4 ug of total RNA were further treated with 1microliter of TURBO DNase I (Ambion, Cat# AM1907) in a 25microliter reaction, according to manufacturer's instructions. RNA quality was assessed before libraries by running all total RNA samples with Bioanalyser RNA 6000 Nano assay (Agilent, cat# 5067-1511) and determining the RNA integrity number (RIN), which was above 7.30 for all samples. 1ug of DNAse-treated total RNA was used for total RNA library production with TruSeq Stranded Total RNA Sample Preparation Kit (Illumina, Cat# RS-122-2201). Library quality was determined with Bioanalyser High Sensitivity DNA assay (Agilent, cat# 5067-4626). Total RNA libraries were sequenced paired-end using using Illumina Sequencing Technology by an Illumina HiSeq2000 following the manufacturer's instructions. Details were gently provided by AM Fenandes.

2.3 Flavopiridol treatment

Flavopiridol datasets were generated by KJ Morris and are published it in her thesis: "Interplay between Polycomb repression and RNA Polymerase II in Embryonic Stem Cells" (Morris, 2012). In Brief, to inhibit Cdk9 OS-25 cells were inhibited with Flavopiridol (10 uM; a kind gift from Sanofi-Aventis, provided by Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute,

Bethesda, MD). for 1h or DMSO for control. Flavopiridol treatment was restricted to 1h to avoid secondary effects on RNAPIIS5p and total RNAPII, as shown in Stock (Stock et al., 2007). Nascent RNA was then extracted by the cells and sequenced.

2.4 Exosome knock down

Exosome si KD was performed by KJ Morris (unpublished). In brief, OS25 cells were transfected with siRNA targeting Exosc3, a catalytic subunit of the Exosome machinery, control siRNA (negative control). Transfection of OS25-ESC was performed with 125 pmole siRNA per 5x105 cells for 48h. Cells were harvested after 48h, RNA was extracted with TRIzol and treated with DNAseI and library were prepared for sequencing as explained on the previous paragraph and in accordance to manufacturer instructions. Untreated cell were also sequenced, however for the differential analyses negative control datasets were considered for comparison.

2.5 RNA-seq datasets processing

Sequenced reads that passed Illumina quality control filters were aligned to the mouse genome annotation (assembly mm10).

Total RNA-seq libraries were mapped with using STAR 2.5(Dobin et al., 2013) with standard options.

Total RNA-seq libraries after exosome KD, negative control and untreated cells were mapped with TopHat v2.0.13 (Trapnell et al., 2012b) were mapped with the options --no-novel-juncs --library-type fr-firststrand

RNA-ChIP-seq libraries in OS25, 46C, treated with Flavopiridol and the control DMSO were mapped with TopHat v2.0.13(Trapnell et al., 2012b) were mapped with the options -r -40 --mate-std-dev 50 --library-type fr-firststrand. RNAChIP in 46C RNAChIP datasets were mapped by RA Beagrie with the same parameters.

Table 2.1: unpublished RNA-seq data mapped in this study

RNA-seq dataset	# sequenced reads	# mapped reads	% mapped reads
RNAChIP (Nascent RNA) (OS25)	243797996	143379774	66.5
RNAChIP (Nascent RNA) (Flavopiridol)	554339578	306774802	55.3
RNAChIP (Nascent RNA) (DMSO)	386142987	183965570	47.6
Total RNA Day 0	163901210	154186012	94.1
Total RNA Day 1	143712874	132245692	92
Total RNA Day 3	154946626	146937742	93.9
Total RNA Day 16	193519812	183247723	94.7
Total RNA Day 30	202503600	193764526	95.7
Total RNA (Exosome si)	63292840	57889016	91.5
Total RNA (Exosome untreated)	72288287	67287013	93.1
Total RNA (Exosome negative control)	62555012	57688806	92.2
mRNA (OS25)	165227729	127030444	75.2

2.6 ChIP-seq dataset handling

Published ChIP-seq data were downloaded from the GEO repository via direct link or using the SRAtoolkit.

ChIP-seq re-mapped and processed by me for the work in the current thesis are: Nanog (Whyte et al., 2013), Sox2(Whyte et al., 2013), Oct4(Whyte et al., 2013), Med1(Whyte et al., 2013), Cdk9(Whyte et al., 2013), H3K4me1(Ferrari et al., 2014), H3K4me3(Ferrari et al., 2014), H3K27ac(Ferrari et al., 2014), H3K122ac(Pradeepa et al., 2016), H3K64ac(Pradeepa et al., 2016), P300(Creyghton et al., 2010), Med12(Rahl et al., 2010), Brd4(Rahl et al., 2010), Nipbl(Rahl et al., 2010), Caph2(Dowen JM, 2018), Cdk8 (Young, unpublished, from http://younglab.wi.mit.edu/datadownload.htm), Cdk7 (Young, unpublished, from http://younglab.wi.mit.edu/datadownload.htm), Spt5(Rahl et al., 2010), Nelfa(Rahl et al., 2010), total RNAPII wt(Tee et al., 2014), RNAPIIS5p wt(Tee et al., 2014), total RNAPII Erk knock out(Tee et al., 2014), RNAPIIS5p Erk knock out(Tee et al., 2014), input control wt(Tee et al., 2014). ChIP-seq and datasets re-mapped and processed by Dr Elena Torlai Triglia: H3K27me3(Mikkelsen et al., 2008), H3K27me3(Ferrai et al., 2017), RNAPIIS5p(Brookes et al., 2012; Ferrai et al., 2017), RNAPIIS7p(Brookes et al., 2012; Ferrai et al., 2017), RNAPIIS2p(Brookes et al., 2012; Ferrai et al., 2017), RNAPIIS2u(Brookes et al., 2012), Smad1 (Chen et al., 2008), Stat3 (Chen et al., 2008), Essrb (Chen et al., 2008), Klf4 (Chen et al., 2008), cMyc (Chen et al., 2008), nMyc (Chen et al., 2008), E2f1 (Chen et al., 2008), H3K36me3(Brookes et al., 2012; Mikkelsen et al., 2008), CTCF(Dowen JM, 2018), Dis3 (AM Fernandes unpublished), Erk2(Tee et al., 2014) , Utf1(Jia et al., 2012). ChIP-seq and datasets re-mapped and processed by Dr Alexander Kukalev: H3K9ac (Consortium, 2012), H3K9me3 (Thibodeau et al., 2017).

Quality control (QC) checks were performed on sequencing data (.fastq files) prior to further processing using FastQC software (Andrews, 2010) and trimming was performed with Flexbar (Roehr et al., 2018) when: adaptor sequences where found present in the sequencing data; per base sequence quality was below 20. Sequenced reads that passed quality check were aligned to the mouse genome annotation (assembly mm10/mm9) using Bowtie2(Langmead and Salzberg, 2012) with default parameters. For each bam file generated after mapping a corresponding indexed .bai file and a .bed were generated with Samtools software (Li et al., 2009). Duplicate reads (identical reads, aligned to the same genomic location) occurring more often than a threshold were removed. The threshold was computed, for each dataset, as the 95th percentile of the frequency distribution of the reads, with an in-house script originally coded by Dr Ines de Santiago and modified by Dr Tiago Rito.

2.7 Bedgraph and bigwig generation

To visualise and analyse RNA-seq and ChIP-seq data, datasets were further transformed. The bedGraph and the bigwig file formats were generated to upload and visualisation on the UCSC genome browser or the IGV browser. These two file formats permit to visualise continuous data on the genome browser to inspect coverage and density of reads. Bigwig was preferred to bedgraph for upload as it is of smaller size and faster to load. Bedgraphs were generateD from bed files through the genome coverage bed command of the BEDTools suite (Quinlan and Hall, 2010). Bigwigs were generated from bedgraph files with the bedtobedgraph command. Data set were loaded on a server to create a permanent link for UCSC genome browser. Reads are visualised always showing the 0 and whit a smoothing window of 2. IGV desktop software(Robinson et al., 2011) was used with the same settings. All RNAPII datasets and H3K27me3_day0 tracks were loaded on the genome browser by Dr Elena Torlai Triglia.

2.8 Expression levels calculation TPM-FPKM and data visualisation

To calculate the transcription levels coverage of RNA-seq data at regions of interest was calculated with multiBamCov of the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) using as input bam files from the RNA-seq data and bed files of the regions of interest. For exon expression, gene file listing all the exons were used and the total number of reads per genes was calculated with a custom script in python. Resulting files were loaded in R and Transcripts per Kilobase per Million (TPMs) and Fragments Per Kilobase Million (FPKMs) were calculated per datasets. log(TPM/FPKM) + 0.001 are showed in the text for clarity.

FPKMs were calculated as:

counts per region/(region length x (total counts/ 10^6))

TPMs were calculated as:

((# counts per regions/region length) x 10^6)/ total # of counts

2.9 ChIP-seq enrichment analysis

ChIP-seq reads enrichment at regions of interest was calculated starting from bed files with intersecBed of the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with the –c parameter. Files were then imported in R and the log average reads count was calculated as follows: the number of reads per regions were divided for the region length. A pseudo count corresponding to the minimum count in the dataset higher than zero divided by ten was added to avoid log(0). Reads are then log transformed.

Formula:

log((# reads per region/region length) + (min(#read in dataset != 0)/10))

Biological replicates were treated separately when present.

2.10 Differential analysis on RNA-seq and ChIP-seq analysis

Differential analysis of ChIP-seq and RNA-seq datasets was performed to compare datasets generated between different treatments or different time points. The analysis was performed with DESeq2 software (Love et al., 2014) in R with count bed files generated as described in 2..7 f or RNA-seq and 2.8 for ChIP-seq data. DESeq2 outputs consist in log2 fold change difference between the two datasets considered and normalised counts, which were used in the plots presented in the current thesis.

2.11 Peak calling on ChIP-seq data

Peaks for all the features presented in this thesis were called with the Bayesian Change-Point (BCP) software (Xing et al., 2012) with histone modification (HM) or transcription factor (TF) settings, depending on the dataset. TF was used for transcription factors; HM was used for histone modifications and RNA Polymerase II. RNAPII and H3K27me3 peaks were called by Dr. Elena Torlai Triglia. BCP peak caller performed poorly on Essrb dataset after visual inspection of both modalities and therefore the Essrb peaks were not further analysed in this work. However, Essrb raw reads were analysed,

2.12 Heatmaps and average plots

Heatmaps of feature enrichment at overlapping enhances in Fig 3.10 were generated using ranked normalised reads (using rank command in R) for comparison between enhancer lists. Hierarchical clustering was calculated based on variance (“Ward.D2” in R) on TFs and histone modification datasets, excluding RNAPII datasets and active transcription marks per subgroup of overlap. Summarised heat map for Whyte enhancers in Fig 4.12 was generated after calculating the average enrichment and ranked normalised per class per factor with the *rank* command in R. Positional heat maps were generated with the Deeptool software v3.0.2 (Ramírez et al., 2018) either starting from the center of the region or averaging the reads per regions +/- #kb of surroundings. Average plots were generated with Deeptools or a self-made R scripts either starting from the center of the region or averaging the reads per regions +/- #kb of surroundings. Biological replicates were treated separately when present.

2.13 Enhancer lists retrieval and manipulation

Published enhancers lists used in the current thesis were downloaded from the relevant publications: Whyte *et al.* 2013 (Whyte et al., 2013), Cruz Molina *et al.* 2017 (Cruz-Molina et al., 2017), Pradeepa *et al.* 2016 (Pradeepa et al., 2016), Arner *et al.* 2015 (Arner et al., 2015), Chen *et al.* 2012 (Chen et al., 2012), Creyghton *et al.* 2010 (Creyghton et al., 2010), Zentner *et al* . 2011 (Zentner et al., 2011). When necessary, liftOver tool from UCSC (Karolchik et al., 2008) was used to change the coordinates to the mouse genomic assembly mm10. The analyses in the current thesis were conducted in the mm10 assembly, unless otherwise specified. Enhancer lists were indexed and annotated for their enhancer classes with self-made bash and python scripts. Arner enhancers were enlarged as to be of a minimum size of 1kb with BEDTools suite v2.17.0 (Quinlan and Hall, 2010), as in the original paper (Arner et al., 2015). Arner enhancers were downloaded from the FAMTOM5 website by Dr Markus Schueler.

2.14 PROMPTs regions generation

PROMPT regions were defined as regions 500/1000bp upstream TSSs of active genes using the UCSC online portal (http://genome.ucsc.edu/cgi-bin/hgTables) selecting the table knownGeneOld6. This file was then filtered to contain only regions upstream active promoters derived from the list of promoter states in Ferrai *et al.* 2017 (Ferrai et alFerrai et al., 2017). Only results on 500bp PROMPTs are presented in this thesis. Results between 500bp and 1kb long PROMPTs are comparable.

2.15 Promoter regions generation

Promoter regions were defied as the region -1000bp, +1000bp around the TSS, as previously (Ferrai et alFerrai et al., 2017). The regions were recovered UCSC online portal (http://genome.ucsc.edu/cgi-bin/hgTables) selecting the table knownGeneOld6. Promoters were annotated for their state using the list of promoter states in Ferrai *et al.* 2017 (Ferrai et alFerrai et al., 2017) and a self-made python script. The states considered in the analysis are: Inactive, Active, Polycomb Repressed (PRC), PRC-Ser5p.

2.16 Distribution of regions across the genome

To show the distribution of regions across different chromosome through the genome, the R package GenomicFeautures package from Bioconductor (Lawrence et al., 2018) was used. The Cruz Molina enhancer list is the only list defined on male cells, therefore identifying a handful of enhancers also on the Y chromosome.

2.17 Distance from the gene analysis for enhancer regions

To calculate the distance of regions of interest promoter regions generated as in 2.15 or gene regions downloaded from the UCSC online portal (http://genome.ucsc.edu/cgi-bin/hgTables) were used. Distance from regions of interest was calculate using closestBed of the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with the following parameters: -d -t first.

2.18 Combination of factors bound at regions of interest

Regions of interested were analysed for combinatorial presence of transcription factors and histone modifications. The combination of factors at regions of interest was calculated as follows: regions of interest were annotate selected features with annotateBed of the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with -counts or -both parameters. As input files the BCP peaks of the selected features and the bed files of the regions of interest were used. The annotated file was then imported in R and the combination of factors plotted with the *UpsetR* package (Conway et al., 2017). If more than one peak overlapped the region it was counted as 1; this was done to allow compatibility with the *UpsetR* package, which analyses at combination of factors as a binary yes/no co-occurrence.

2.19 Co-localisation analysis

To calculate the co-localisation between regions of interest intersectBed from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) was used with the following parameters: -wa –u (-wb). To calculate the regions not co-localising intersectBed was used with the –v parameter. Regions generated in this way were saved as new bed files for enrichment analysis.Venn plot showing co-localisation of regions of interest were generated in R with the *venneuler* package (Wilkinson, 2012).

2.20 TFs binding at co-localising enhancers analysis

In Chapter 3, I analyse the number of TFs bound per region. To calculate the number of TF binding at regions of interest, the regions were annotate for TF BCP peaks with annotateBed from BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with –count option. Generated files were imported in R for data visualisation. Regions of interest were divided in groups based on the number of TFs bound for further analysis. Bins were calculated as: 0 TFs, 1 TF, 2 TFs, 3 TFs, 4 or more TFs.

2.21 RNAPII and CAGE tags occupancy at co-localising enhancers analysis

To calculate the number of RNAPII peaks and CAGE tags binding at overlapping enhancer regions, regions of interest were annotated for BCP peaks of RNAPIIS5p_1, RNAPIIS5p_2, RNAPIIS2u, RNAPIIS7p, RNAPIIS2p, RNAPII-CTDK7me1, RNAPII-CTDK7me3 and CAGE tags with annotateBed of the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with the –count parameter. The resulting file was imported in R and converted in a binary yes: 1 or more peaks overlapping; no: 0 peaks overlapping files for RNAPII and CAGE tags datasets separately.

2.22 RNAPII binding per quartile analysis

To analyse the RNAPII binding depending on the TF enrichment, regions of interest overlapping with one TF bound were analysed. ChIP-seq average counts of the TF of interest per region were calculated and the regions divided for belonging to the quartiles: 0-25%, 25-50%, 50-75%, and 75-100%. Quartiles were calculated with the *summary* function in R. RNAPII presence derived from files created as described in 2.21 was analysed per quartile and plotted as a fraction of regions covered per quartile.

2.23 Correlation plots

Correlations between features at regions of interest were calculated in R with the command *dist* and transformed into a matrix. Correlation plots were generate with the *corrplot* package in R with the *PCA* for clustering option.

2.24 Division of enhancers in extragenic and intragenic

Regions of interest were divided in intragenic and extragenic based on their location. Intragenic regions were defined as: co-localising with an annotated gen from RefSeq (O'Leary et al., 2016), 2kb upstream of an annotated TSS, and co-localising with an RNAPII covering an annotated gene and extending beyond the annotated TES. Extragenic region were defined as outside intragenic regions previously described. The minimum overlap to define a region intragenic was set at 1bp.

2.25 Division of RNAPII peaks in extragenic and intragenic

RNAPII peaks were defined as intragenic and extragenic as defined in 2.24, with the addition of the use of UCSC genes (Casper et al., 2018) to define annotated regions. Regions were also filtered for not co-localising with the ENCODE blacklist (Consortium, 2012).

RNAPII peaks defined across neuronal differentiation were divided in intragenic and extragenic as previously, with the addition that intragenic region comprised co-localising regions with an RNAPII covering an annotated gene and extending beyond the termination site for all time points.

2.26 Calculation of the length of RNAPII peaks protruding from annotated gene termination sites

To define the length of RNAPII peaks protruding over transcription termination sites of annotated genes, RNAPII peaks crossing a transcription end site were selected and cut at the transcription end site location. The length of the peak from the cut to the end not co-localising annotated genes was plotted in R. Files generated for all RNAPII peaks were merged with mergeBed from BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with standard parameters and used to define intragenic regions as described in 2.24 and 2.25.

2.27 Generation of Random regions for enhancer analysis

For every set of region of interest a matched set of random regions were generated. Random regions were generated randomly shuffling the original regions of interest in the genome, to allow fair comparison with shuffleBed from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with the parameters: -noOverlapping –excl <original regions of interest> to restrict their localisation so not to co-localise with the original regions of interest.

Random regions matched to extragenic regions were further divided in intragenic and extragenic with the same procedure applied to their matched regions of interest.

For the analysis of co-localising enhancers in chapter 3 only a set of random regions was produced. For the random regions generated in chapter 4 and in chapter 5 30 different permutations were performed and used separately to gain statistical power in calculation of classification cut-off and in random regions classification were used to calculate error bars. Random regions used to define the classification threshold and classified random regions used as reference were calculated separately. Random regions showed in boxplots are one representative random region.

2.28 Classification of enhancer regions for RNAPII states

To classify extragenic enhancer regions for RNAPII states different step were taken, describe below.

Enrichment of RNAPII modification

Enrichment for every extragenic enhancer region and for matched random regions was calculated with intersectBed command from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) and the option –c. Regions longer than 3kb were divided in 3kb bins and the enrichment was calculated per bin. The largest enrichment across all the bins was assigned to the original region with a sel-made python script using *pandas* package (McKinney, 2018). Random regions used in this step are not used for the annotation analysis below.

Annotation for RNAPII peak

Extragenic enhancer regions and matched random regions were annotated for RNAPII peak co-localisation with annotateBed from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with the –both option. Random regions annotated for RNAPII peaks are not used in the enrichment analysis.

Classification as positive or negative for a RNAPII modification

To calculate the 5% false positive threshold over random regions enrichment, the enrichment per region is divided by its length and a log transformation is performed. 5% top enrichment is then calculated in R with the *quantile* function. The cut-off value is then used to classify as positive or negative for a RNAPII modification extragenic enhancer regions: regions with a length normalised enrichment (log) above the threshold are considered positive, while the other are negative. These regions constitute the "liberal" list. Extragenic enhancer regions annotated as positive for the RNAPII modifications were classified as positive and negative separately and constitute the "conservative" list.

Definition of RNAPII states via modification combinations

After generation the classification for single RNAPII modifications, files were merged and extragenic enhancer regions are classified per RNAPII activation state. Combination of RNAPII modification presence at the same region generated in the previous step was analysed with a self-made script in R and python. Classification was performed also on the random region datasets annotated for BCP peaks, for comparison.

2.29 Density plots

Density plot for enrichment of features in regions of interest are generated in R with the *density* function. Densities of subset of regions, such as BCP positive regions compared to all regions, are normalised for the maximum height of all regions in R in the following way:

d_BCP$y <- d_BCP$y x (# regions BCP+ /# total regions)

Where: d_BCP is the density of the enrichment of the investigated feature at BCP positive regions, and d_BCP$y is the height of the density curve.

2.30 Gene Ontology analysis of extragenic Whyte enhancers

Gene ontology analysis of Whyte extragenic enhancers positive and negative for RNAPII occupancy was performed with the online software GREAT (McLean et al., 2010) with the standard parameters and the whole genome as background, as previously (Cruz-Molina et al., 2017). GREAT analysis on extragenic RNAPII regions failed the yield any meaningful result because these regions are very far from genes. These analyses were therefore no included in the results.

2.31 Ranking enrichment analysis for super enhancers identification

The analysis of ranking enrichment for super enhancer identification was performed in R as in (Whyte et al., 2013). In brief, enrichment of selected features was calculated at all Whyte enhancers and the regions were ranked for the amount of enrichment from the lowest to the highest. Regions were then plotted on the x-axis for their enrichment ranking and on the y-axis for the amount of enrichment.

2.32 Definition of extragenic RNAPII regions

To define extragenic RNAPII regions, RNAPIIS5p, RNAPIIS7p and RNAPIIS2u peaks from Brookes *et al.*(Brookes et al., 2012) and Ferrai *et al. (Ferrai et alFerrai et al., 2017)* were first divided in extragenic and intragenic as in 2.25. Classification was then performed similar to 2.28 over matched random regions for RNAPIIS5p enrichment at extragenic RNAPIIS5p peaks; RNAPIIS7p enrichment at extragenic RNAPIIS7p peaks; RNAPIIS2u enrichment at extragenic RNAPIIS2u peaks. Extragenic RNAPII regions were selected among the ones with an enrichment above the calculated cut-off. This procedure was applied to select high confidence regions and filter out regions with poor enrichments.

Extragenic RNAPIIS5p regions during neuronal differentiation were defined as described above per time point. Regions were then merged in a single, non-redundant file for further use. Merging was performed as follows: regions were concatenated in a unique file in bash using the *cat* command, sorted for chromosome and start site and merge to avoid redundancy with mergeBed from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with standard parameters.

2.33 Classification of extragenic RNAPII regions for RNAPII states

The classification of extragenic RNAPII regions for RNAPII enhancer state was conducted similarly to the extragenic enhancer classification explained in 2.28. They differ in the fact that extragenic RNAPII regions were classified only as in the "liberal" approach. The "liberal" approach considers a region positive for a RNAPII modification if the length normalised enrichment of the region is above the top 5% enrichment at matched random regions and it is also co-localising with a BCP peak of the feature of interest. RNAPII states are then defined as in 2.28

Extragenic RNAPII regions during differentiation were classified for RNAPII and H3K27me3 as explained above. RNAPII states are then defined per time point as in 2.28.

2.34 Comparison between extragenic RNAPII datasets

Comparison between modifications

To compare different extragenic RNAPII modification regions, regions were first analysed for their co-localisation similar to 2.19. Different overlapping groups were divided and annotated for TFs binding and histone modifications and plotted as in 2.20 and 2.21.

Comparison between mESC clones

To compare extragenic RNAPII regions derived by different mESC clones, regions were first analysed for their co-localisation similar to 2.19. Different overlapping groups were divided and annotated for TFs binding and histone modifications and plotted as in 2.20 and 2.21. Density were calculated and plotted as in 2.29.

2.35 Analysis of co-regulated RNAPII extragenic regions

Regions sensitive for transcription perturbation were analysed in R. Differential enrichment per regions ws calculated as in 2.10 for every feature of interest. Log2FoldChange from DESeq2 results was used to confront sensitivity of regions for different treatments.

2.36 Generation of heat map of waves of extragenic RNAPII states during neuronal differentiation

Extragenic RNAPII regions per time point classified for different states as in 2.33 were plotted order for RNAPII activation state (H3K27me3, H3K27me3-RNAPIIS5p, H3K27me3-RNAPIIS5p-S7p, RNAPIIS5p, RNAPIIS5p-S7p, RNAPIIS7p, Inactive) and time point (Day 0, Day 1, Day 3, Day 16, Day 30) and plotted. Waves of extragenic RNAPII states were also plotted in a simplified way grouping regions in more general states: Polycomb: if a region is

positive for Polycomb, with or whitout RNAPII; Active: if a region if positive for RNAPIIS5p and/or RNAPIIS7p and not Polycomb; Inactive: if the regions is negative for all the marks.

2.37 Transition of RNAPII states during differentiation

To calculate the percentage of transitions between days of differentiation, all transitions between two days were divided by the total number of regions to obtain a percentage and plotted.

2.38 VISTA tested regions analysis

VISTA tested regions retrieval

VISTA tested regions were downloaded from the VISTA enhancer database (Visel et al., 2007) as FASTA files and transformed into a bed file with custom bash script. Only mouse enhancers were downloaded in mm9 forma. The VISTA tested regions were: positive in brain, positive in heart, positive in limb, negative. The four lists were liftOver to mm10 with the online liftOver software, annotated for their state on the VISTA enhancer database and a unique concatenated file was generated to confront with extragenic RNAPII regions.

Overlap between VISTA regions and extragenic RNAPII regions

The overlap between VISTA tested regions and extragenic RNAPII regions was performed with intersectBed command from the BEDTools suite v2.17.0 (Quinlan and Hall, 2010) with parameters: -wa -wb. Extragenic RNAPII regions used for the comparison were: extragenic RNAPII active at Day 16 and Day 30; active at Day 16 and inactive at Day 30; inactive at Day 16 and active at Day 30; Polycomb repressed at Day 16 and at Day 30 (data not shown). The same analysis was performed between Day 0 and Day 1 and Day 1 and Day 3. Day 0 and Day 1 analysis is not shown in the text of the current thesis, as the VISTA tested regions don't cover enhancers active in early development and the majority of extragenic RNAPII regions were negative at both time points.

2.39 Custom scripts and plot generation

All the plots showed in the current thesis were generated in R with R plotting functions or ggplot2 package (Wickham, 2009), unless otherwise specified. All plotting scripts were custom made, unless specified.

3. Understanding enhancer classifications

3.1 Introduction

Transcription factor (TF) binding at regulatory regions is associated with enhancer activity (Palstra and Grosveld, 2012): specific TFs bind a consensus region on the genome and regulate enhancers. The regulation can be towards activation, but also towards repression. Interestingly, combinations of TFs can form complexes known as enhanceosomes, composed of different TFs and can achieve different regulatory outcomes (Merika and Thanos, 2001). Therefore, the combination of TFs bound at a specific enhancer can be informative of their state. However, our current understanding of TF binding and its relation with enhancer activation states is lacking.

Enhancers exist in different states (active, poised, primed, etc) (Calo and Wysocka, 2013). The study of these different states has helped understand how, for example, poised enhancers keep genes in a repressed state in stem cells and ready for activation during neuronal differentiation (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). However, single studies concentrate on particular classes of interest, lacking a global comparison between different classes of enhancers. Recently, Benton and colleagues compared active enhancers defined with different strategies. The authors showed that regions identified as enhancers by multiple approaches were not more likely to act as regulatory regions than those identified by a single approach; the former were not more often found to be active in transgene assays or in the VISTA enhancer database (Visel et al., 2007) than putative enhancers uniquely found by one approach (Benton et al., 2017).

The combinatorial nature of TF binding to enhancers, the different activation states, and the differences between the approaches used to find enhancers leave open the question on how all these features relate with each other. What kinds of regions are identified when choosing one approach over another? Do they have similar or different properties?

3.2 Aim of the chapter

The aim of the current chapter is to understand differences and similarities between regulatory regions identified through different enhancer classification strategies and definitions, and to investigate their characteristics in general, but especially with respect to TF binding (Scheme of the chapter Fig 3.1).

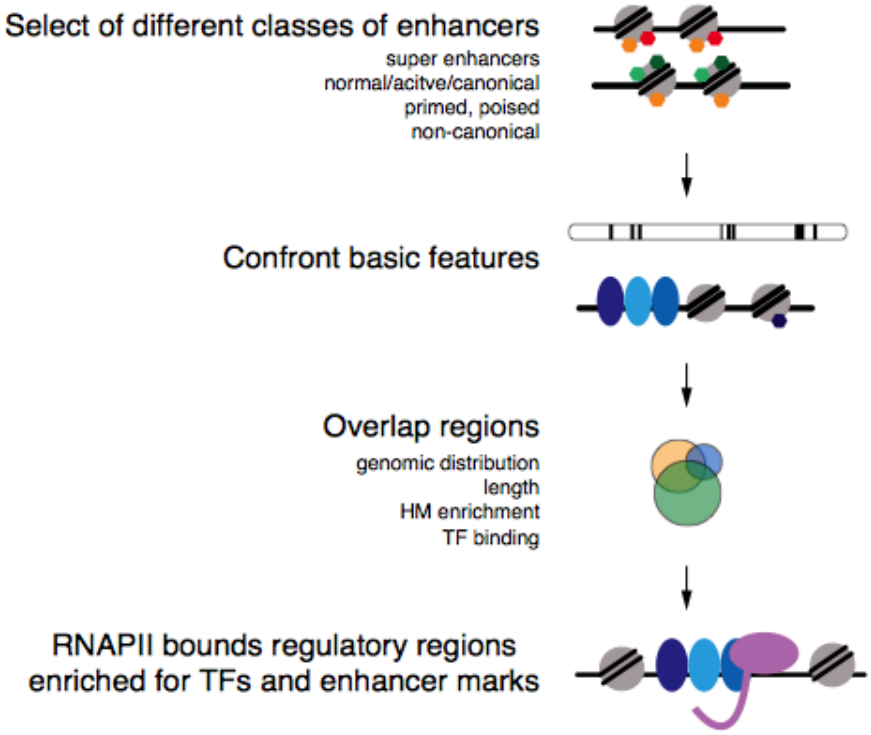


Fig 3.1: Overview of Chapter 3. Enhancers classified with different strategies were analysed for feature enrichment and Transcription Factors binding. We find that RNAPII binding co-occurs with Transcription Factor binding at regulatory regions.

I started by analysing three different published lists of enhancers that define five different classes of enhancers in mouse embryonic stem cells (mESC): super enhancers, normal/active/canonical enhancers, primed enhancers, poised enhancers, and non-canonical enhancers (Fig 3.2).

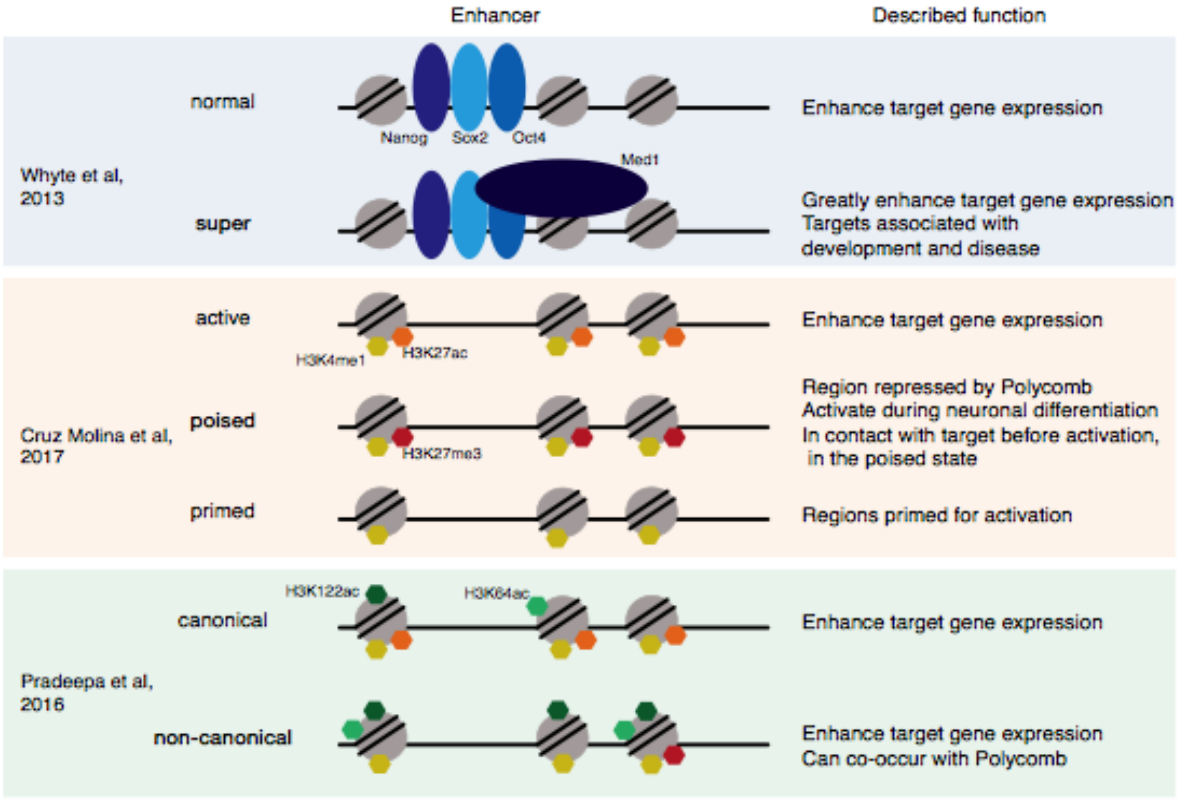


Fig 3.2: Overview of the enhancer classes analysed in this chapter and their described function. In bold the main class explored in the original publication. To enhance clarity, the schematic doesn't reflect the exact proportions and position of protein and histone modifications.

The work presented in the current thesis describes to what extent different enhancer identification approaches find different regulatory regions characterized by diverse sets of features. Combinations of TF binding at candidate enhancers co-occurs with RNAPII occupancy, suggesting RNAPII as a possible new feature to define enhancers at extragenic regions, and their activation state.

3.3 Contribution disclosure

Elena Torlai Triglia processed and calculated the peaks for all the RNAPII modifications datasets presented in the current chapter. Elena Torlai Triglia and Alexander Kukalev processed some of the datasets used in the current chapter, when specified. Markus Schueler downloaded published CAGE tags regions and converted them from mm9 to mm10.

3.4 Results

3.4.1 Choice of published enhancer lists

To understand how different classes of enhancers compare to each other, I obtained three published lists of enhancers (Whyte *et al.* 2013 (Whyte et al., 2013), Pradeepa *et al.* 2016 (Pradeepa et al., 2016), Cruz Molina *et al.* 2017 (Cruz-Molina et al., 2017)), classified in mouse

embryonic stem cells (mESCs) based on the presence different markers; these lists will be referred throughout as Whyte, Cruz Molina and Pradeepa. The three lists cover the five main classes of enhancers described in the literature: super enhancers, normal/active/canonical enhancers, primed enhancers, poised enhancers, and non-canonical enhancers. Whyte, Cruz-Molina and Pradeepa classifications are based on the presence of different features to identify candidate enhancers outside promoter regions: presence of histone modifications (HM) and transcription factor (TF) binding (Table 3.1). The Whyte list catalogues regulatory regions as *normal* enhancers if bound simultaneously by Nanog, Sox2 and Oct4 and clustered within 12.5kb, and as *super* enhancers (SE), the regions among normal enhancers that are highly enriched for Med1 (Whyte et al., 2013). The Cruz Molina list uses an approach based on histone-modification enrichment to classify: *active* enhancers, when enriched for H3K27ac and H3K4me1; *primed* enhancers when enriched only for H3K4me1; and *poised* enhancers when enriched for H3K27me3 and H3K4me1. Finally, the Pradeepa list also uses a histone-modification based approach: *canonical* enhancers are marked by H3K27ac and H3K4me1, while *non-canonical* enhancers are marked by H3K4me1, H3K122ac, but not by H3K27ac. Canonical active enhancers of the Pradeepa list can also be occupied by H3K122ac (Pradeepa et al., 2016).

Table 3.1: List of enhancer lists considered in this chapter divided by enhancer class. Publication, number of enhancers, mESC clone and marks used for the classification are indicated.

Publication	Class	Number of enhancers in class	Features used for the classification	mESC clone
Whte *et al.*, 2013	Normal enhancer	8563	Nanog+ Sox2+ Oct4+	V6.5
Whyte *et al.*, 2013	Super enhancer	231	Nanog+ Sox2+ Oct4+ Med1+	V6.5
Cruz Molina *et al.*, 2017	Active	12142	H3K4me1+ H3K27ac+ H3K27me3-	E14
Cruz Molina *et al.*, 2017	Poised	1015	H3K4me1+ H3K27ac- H3K27me3+	E14
Cruz Molina *et al.*, 2017	Primed	19723	H3K4me1+ H3K27ac- H3K27me3-	E14
Pradeepa *et al.*, 2016	Canonical	23149	H3K4me1+ H3K27ac+	!"#$
Pradeepa *et al.*, 2016	Non Canonical	9340	H3K4me1+ H3K27ac- H3K122ac+	!"#

The choice of these three lists enabled me to investigate classes of enhancers in different states, from the less active (primed enhancers), to the active enhancers and the super enhancers, but also the poised and the non-canonical enhancers. More enhancer lists were published on mESC during the course of this PhD project: in the next chapter some analysis on the Chen *et al.* 2012 (Chen et al., 2012) and Arner *et al.* 2015 (Arner et al., 2015) lists are presented. However a

profound analysis on these lists or other recently published ones have not been considered here for time constrain reasons.

3.4.2 Features used to characterise enhancer classes

To study the features associated with the different enhancer classes considered, ChIP-seq datasets were downloaded as raw data and re-processed for consistent analysis (Table 3.2). Available occupancy sites were avoided for a number of reasons: first, I wanted to process the datasets in the same, reproducible way and with the same quality standards; second, not all the features analysed provided occupancy sites; third, available occupancy sites differed in their processing between publications. Some datasets, such as CTCF and Nanog from ENCODE, where found of poor quality due to high noise and not considered in the analyses presented here. After data re-processing, peaks were called for all datasets using the BCP software (Xing et al., 2012). BCP software performs well on narrow and broad peaks and especially well in mixed peaks such as RNAPII peaks (Harmanci et al., 2014; Thomas et al., 2017). Of note, BCP software, as well as MACS2 software (Zhang et al., 2008), performed badly with the Essrb datasets (data not shown) and therefore peaks of Essrb were not used in the analysis, whereas Essrb ChIP-seq raw reads were used.

Table 3.2: Published ChIP-seq datasets used in this work. * Re-mapped and processes by Elena Torlai Triglia; † Re-mapped and processed by Alexander Kukalev. All peaks of the datasets in this list were computed by Giulia Caglio, except H3K27me3 day0 and RNAPII datasets, computed by Elena Torlai Triglia.

Dataset	mESC clone	Publication	Type	GEO	ab clone
Nanog	V65	Whyte *et al.*, 2013	trancription factor	GSM1082342	Bethyl A300-397A
Sox2	V65	Whyte *et al.*, 2013	trancription factor	GSM1082341	R&D Systems MAB2018
Oct4	V65	Whyte *et al.*, 2013	trancription factor	GSM1082340	sc-8628X
H3K4me1	E36	Ferrari *et al.*, 2014	histone modification	GSM970225	abc895
H3K27ac	E36	Ferrari et*et al.*, 2014	histone modification	GSM970221	ab4729
H3K27me3_1*	V65	Mikkelsen *et al.*, 2008	histone modification	GSM307619	Upstate 07-449
H3K27me3_2*	46C	Ferrai *et al.*, 2017	histone modification	GSM2474113	Millipore, # 07-449
H3K122ac	46C	Pradeepa *et al.*, , 2016	histone modification	GSM2054689	Tropberger et al 2013
H3K64ac	46C	Pradeepa *et al.*, 2016	histone modification	GSM2054691	Di Cerbo et al 2014
Smad1*	E14	Chen *et al.*, 2008	trancription factor	GSM288348	sc-7965
Stat3*	E14	Chen *et al.*, 2008	trancription factor	GSM288353	sc-482
Essrb*	E14	Chen *et al.*, 2008	trancription factor	GSM288355	custome made ab
Klf4*	E14	Chen e*et al.*, 2008	trancription factor	GSM288354	custome made ab
cMyc*	E14	Chen *et al.*, 2008	trancription factor	GSM288356	custome made ab
nMyc*	E14	Chen *et al.*, 2008	trancription factor	GSM288357	custome made ab
E2f1*	E14	Chen *et al.*, 2008	trancription factor	GSM288349	Upstate 05-379
H3K9ac†	E14	ENCODE	histone modification	GSM1000123	ab4441
H3K9me3†	46C	ENCODE	histone modification	GSM1003751	ab8898
H3K4me3	E36	Ferrari *et al.*, 2014	histone modification	GSM970227	Active Motif 39159
H3K36me3_1*	V65	Mikkelsen *et al.*, 2008	histone modification	GSM307606	ab9050
H3K36me3_2*	OS25	Brookes *et al.*, 2012	histone modification	GSM850472	13C9
RNAPII-S2u	OS25	Brookes *et al.*, 2012	RNAPII modification	GSM850469	8WG16
RNAPII-S5p_1	OS25	Brookes *et al.*, 2012	RNAPII modification	GSM850467	CTD4H8
RNAPII-S5p_2	46C	Ferrai *et al.*, 2012	RNAPII modification	GSM2474111	CTD4H8
RNAPII-S7p	OS25	Brookes *et al.*, 2012	RNAPII modification	GSM850468	4E12
RNAPII-S2p	OS25	Brookes *et al.*, 2012	RNAPII modification	GSM850470	H5
RNAPIIK7me1	OS25	Dias *et al.*, 2015	RNAPII modification	GSM1874007	K7me1 3D3
RNAPIIK7me2	OS25	Dias *et al.*, 2015	RNAPII modification	GSM1874008	K7me2 19B4
P300	V65	Creyghton *et al.*, 2010	co-factor	GSM594600	sc-585
Med12	V65	Rahl *et al.*, 2012	co-factor	GSM560354	Bethyl A300-774A
Brd4	V65	Rahl *et al.*, 2012	co-factor	GSM937540	Bethyl A301-985A
CTCF*	E14	Chen *et al.*, 2008	chromatin topology	GSM288351	Upstate 07-729
Nipbl	V65	Rahl *et al.*, 2012	chromatin topology	GSM560349	Bethyl A301-779A
Caph2	V65	Dowen *et al.*, 2013	chromatin topology	GSM824836	Bethyl A302-275A
Med1	ZHBTc4	Whyte *et al.*, 2013	co-factor	GSM1038259	Med1/TRAP220
Control	OS25	Brookes *et al.*, 2012	mock IP	GSM850473	-

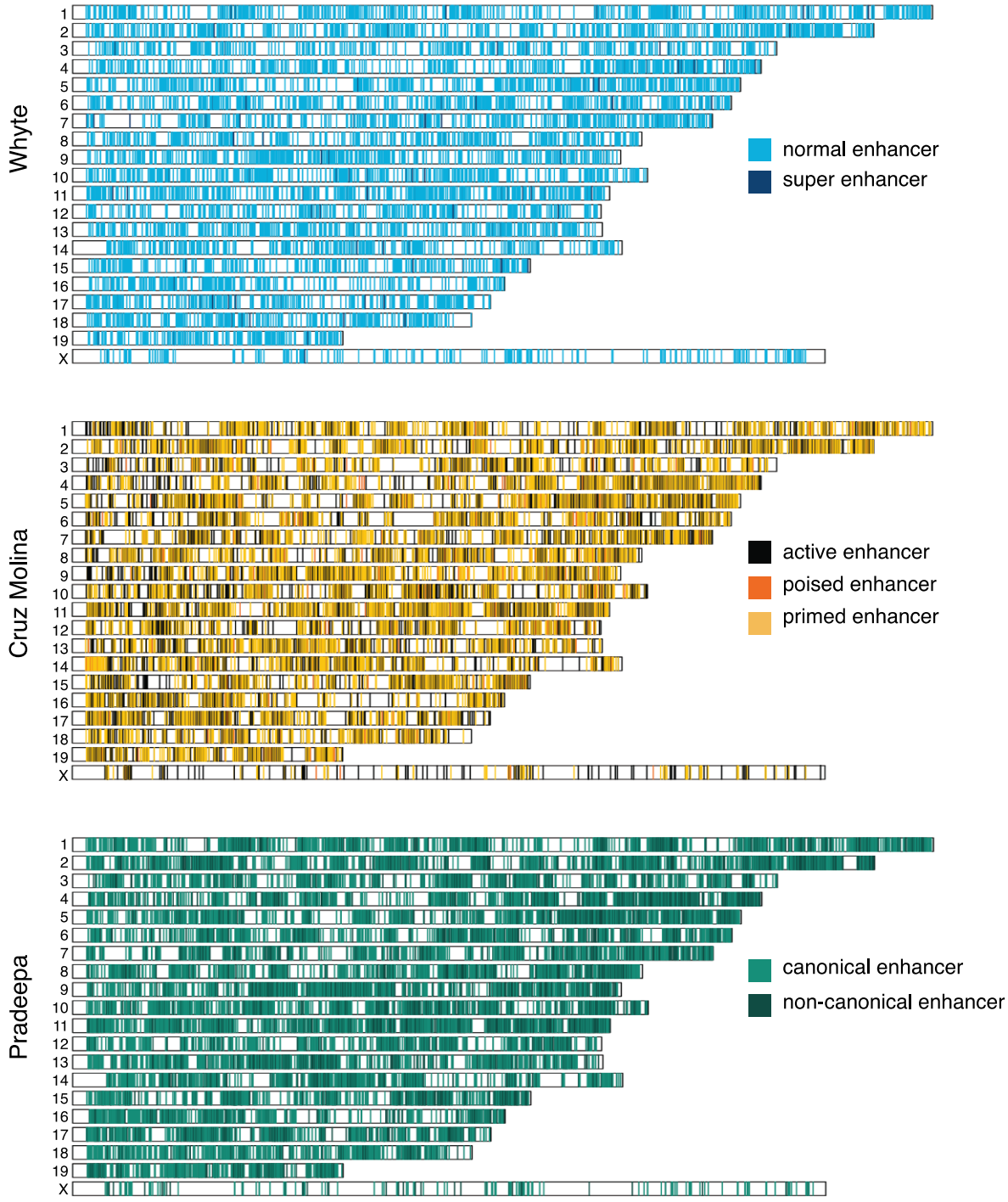
Briefly, I have analysed the classifier marks used to define the different enhancers (Nanog, Sox2, Oct4, H3K4me1, H3K27ac, H3K27me3, H3K122ac, H3K64ac), known TFs (Smad1, Stat3, Essrb, Klf4, cMyc, nMyc, E2f1), histone modifications that mark active chromatin (H3K9ac), inactive chromatin (H3K9me3), promoters (H3K4me3), transcribed regions (H3K36me3), RNAPII modifications (Unphosphorylated Ser2 (8WG16ab), RNAPII Ser5p, Ser7p, Ser2p, CTDK7me1, CTDK7me2), co-factors (P300, Med12, Brd4), and proteins involved in genome architecture (CTCF, Nipbl, Caph2, Med1). I have also analysed Mock IP controls. In some cases, I considered more than one dataset per feature (e.g. H3K27me3). Taken together, the features considered here define different states of chromatin: heterochromatin, promoters, transcribed coding regions, putative enhancers, and insulators or looping domains.

3.4.3 Enhancer classes distribution across the genome

To find the similarities and differences across enhancer classes and lists, I started by mapping the position of the different enhancers onto their chromosome coordinates. The distribution of enhancers across the genome is quite uniform between datasets and enhancer classes (Fig 3.3 whole genome a – chromosome 2 zoom-in b).

Figure 3.3

a Distribution per chromosome



b distribution chromosome 2

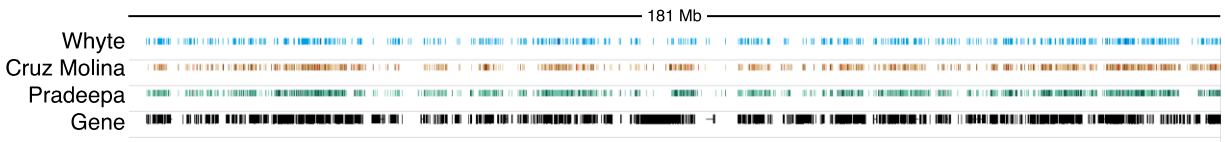


Fig 3.3: Enhancers identified in different lists are widespread across the genome. a) Location of Whyte's, Cruz Molina's, and Pradeepa's enhancers across the genome. Chromosomes are ordered by number (specified on the left). Lists and classes of enhancers are color-coded. Whyte enhancers: normal – light blue, super – dark blue; Cruz Molina enhancers: active – brown, poised – orange, primed – gold; Pradeepa enhancers: canonical – light green; non-canonical – dark green. b) Location of Whyte, Cruz Molina, and Pradeepa enhancers and Refseq genes across chromosome 2. Blue lines indicate Whyte enhancers, golden lines indicate Cruz Molina enhancers, green lines indicate Pradeepa enhancers, and black line indicate Refseq Genes. Image generated with the IGV software.

However, closer inspection of specific gene loci shows a complex scenario Fig 3.4a). For example, SEs can overlap with canonical, non-canonical, active, and primed enhancers (see Nanog locus). Interestingly, poised enhancers, which contact repressed target genes in mESC, can reside in SE regions (see Tbx3 locus). Additional examples of enhancer distributions at specific genomic regions can be found for loci Sox2, Foxd3, Gm5607 and Lif, in Appendix Fig 3.A1. Taken together, these examples clearly show that enhancer lists are ambiguous, with different genomic regions being classified in varied ways depending on the subset of features considered.

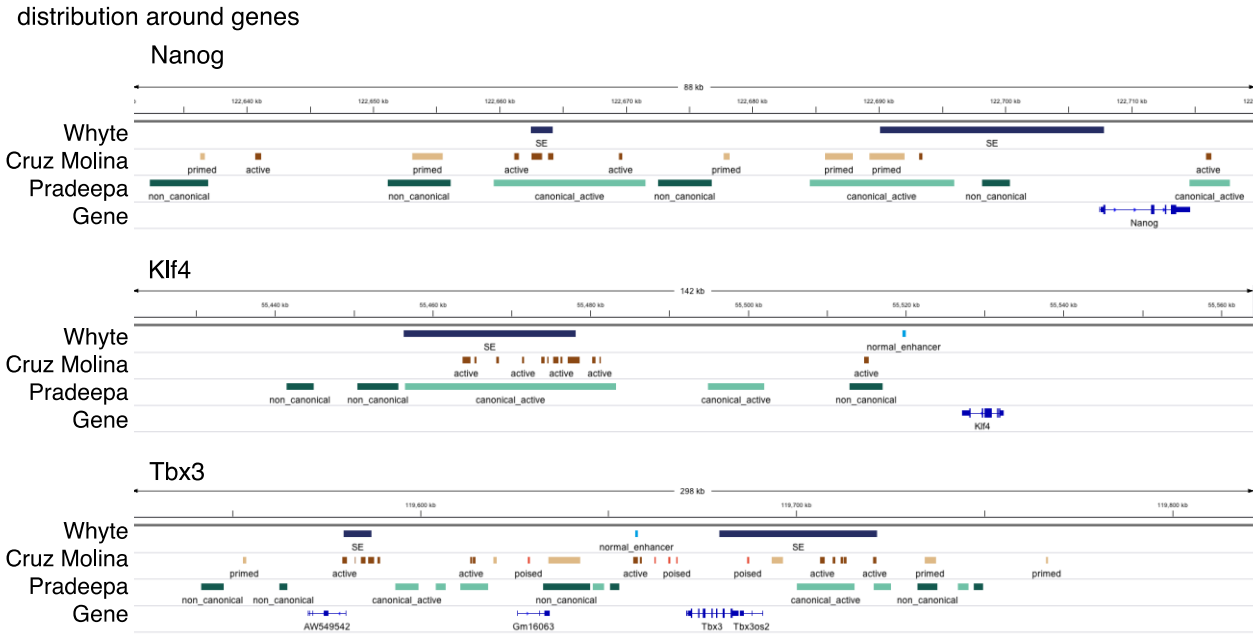


Fig 3 4: Regulatory landscape at gene loci. a) Enhancer classes locations at 3 gene loci. Whyte, Cruz Molina,Pradeepa enhancers and RefSeq genes are shown as coloured boxes. Classes of enhancers are indicated below the enhancers. Whyte enhancers: normal – light blue, super – dark blue; Cruz Molina enhancers: active – brown, poised – orange, primed – gold; Pradeepa enhancers: canonical – light green; non-canonical – dark green. Nanog locus was chosen because of its relevance in mESC. Klf4 locus was investigate as SE targets in previous publications 8. Tbx3 locus was investigated as non-canonical enhancer's targets 9. Images generated via the IGV software.

3.4.4 Enhancer features differ between classification lists

To explore whether candidate regulatory regions identified in each list have different properties, I first examined their length and distance to the most proximal gene. These two measures could be informative of enhancer function and mechanism: distal enhancers could be involved in chromatin looping, or long enhancers could act as TF docks.

SEs tend to be longer than normal enhancers in the Whyte list, with medians 8667bp and 703bp, respectively (Fig 3.5a). Pradeepa's canonical and non-canonical enhancers are in general longer (median 4000bp and 3000bp, respectively), while the Cruz Molina enhancers are shorter (median 431bp, 353bp and 1265bp, for active, poised, and primed enhancers). These differences can influence the TF binding and histone modification occupancy across different lists of enhancers.

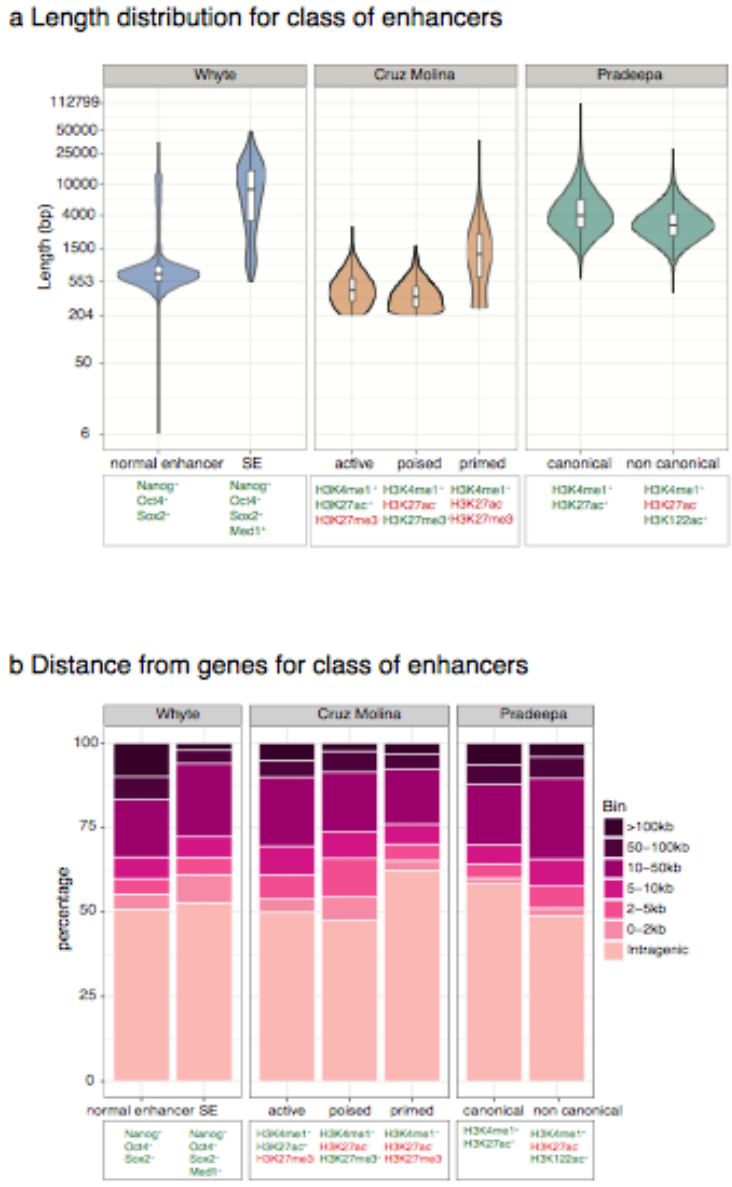


Fig 3 5: General features of enhancer classes. a) Distribution of length of enhancers. Whyte, Cruz Molina and Pradeepa enhancers are shown, divided in enhancer classes. The box on the bottom specifies the marks used for each class definition: green – mark present; red – mark absent. b) Distribution of enhancer classes' distance from nearest annotate TSS-TES in bp. Intragenic: the enhancer region is inside an annotated gene. Whyte, Cruz Molina and Pradeepa enhancers are shown, divided in enhancer classes. The box on the bottom specifies the marks used for each class definition: green – mark present; red – mark absent.

Next, I measured the distance of enhancers to most proximal gene (Fig 3.5b) to understand whether different classes of enhancers have preferential positions. For all lists, half of the candidate enhancer regions are contained inside a coding region and 10-20% are located at a

distance between 0 to 10kb from the closest transcription start site (TSS) or transcription termination site (TES). For all lists, 10 to 20% of the enhancers lie at more than 50kb of the most proximal gene, with noticeably some at more than 100kb; the Whyte list captures the largest proportion of enhancers in this category (855/8794, ~10%).

In conclusion, these results show that the different approaches to define enhancers are consistent in their genomic location in respect to genes, with ~50% residing inside them, however they also show differences in the features associated with each region. Some enhancers tend to be significantly shorter than others, potentially influencing features occupancy, and this effect is mainly dependent on the approach used to define enhancers.

3.4.5 Binding of transcription factors and structural proteins differ between classes of enhancers

Transcription factors have a fundamental role in enhancer activation and in keeping cell identity: binding of different combinations of these proteins together with transcription co-factors, and chromatin re-modellers can inform about the potential activity of enhancer regions. To explore whether enhancer classes vary in their protein binding and whether any differences are connected to activation states, I analysed the binding of transcription factor, transcription co-factors, and chromatin re-modellers.

In the Whyte list, most regions are bound by Nanog, Oct4, and Sox2 (6748 /8794, 77%), or combinations of at least two of these three factors (93%; Fig 3.6a, for other combinations see Appendix Fig. 3.A2). The few enhancers not classified here as positive for either Nanog, Sox2, or Oct4 (1.2%) are probably due to slight differences in peak calling between the current thesis and the original Whyte paper. Interestingly, Nanog, Sox2, and Oct4 binding is associated with the concomitant binding of proteins involved in chromatin looping, such as Nipbl (cohesin loading factor), at normal enhancers, but especially at SEs, possibly because they are involved in chromatin loops (Hnisz et al., 2013; Whyte et al., 2013) (~30 and ~90% respectively). Other TFs expressed in mESCs, such as nMyc and cMyc, together with Smad1, are not found frequently at Whyte enhancers.

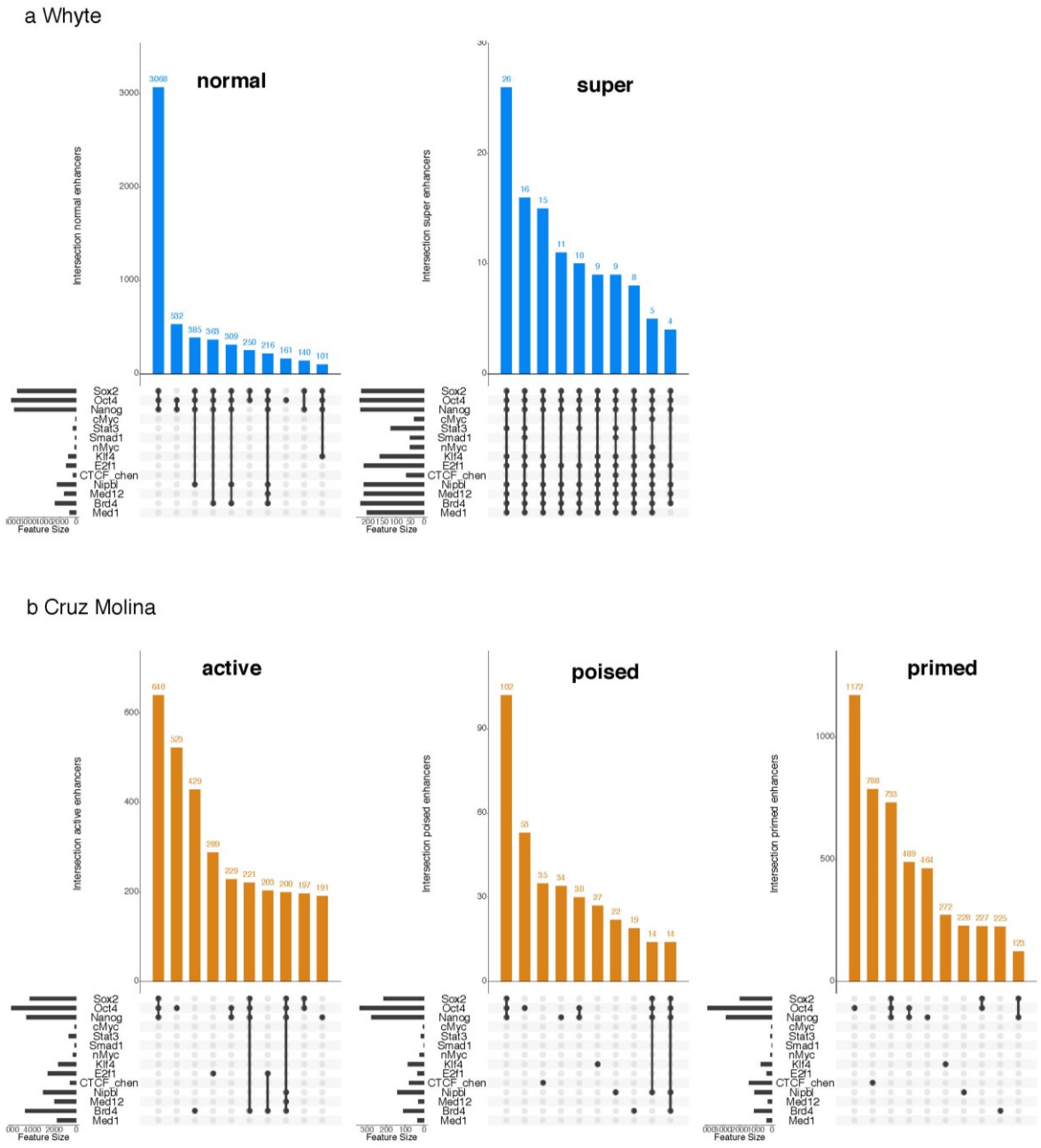


Fig 3.6: Different classes of enhancers have diverse TFs and structural proteins binding. a) Plot showing the binding of different proteins at Whyte enhancers. On the box below: combination of factors. On the y-axes the number of enhancers bound by the combination indicated. On the x axis on the left: total number of enhancers with that specific protein bound. b) Plot showing the binding of different proteins at Cruz Molina enhancers. On the box below: combination of factors. On the y-axes the number of enhancers bound by the combination indicated. On the x axis on the left: total number of enhancers with that specific protein bound.

In the Cruz Molina list, the most represented TFs remain Nanog, Oct4, and/or Sox2 (at least one: ~56% of active, ~40% of poised, ~25% of primed). The most represented combination of TFs at active and poised enhancers is Nanog, Oct4, and Sox2 (25%, 17% respectively), while at primed enhancers Oct4 is mainly found alone (~20%) (Fig 3.6b). CTCF, a transcription factor thought to be important in insulation and chromatin looping, is not preferentially found at active enhancers, and only in a small subset of primed enhancers (7%). Only a small subset of Cruz Molina

enhancers are bound by CTCF (>7%), while Cohesin is bound to Cruz Molina enhancers the more they are in an active state (5% of primed, 13% of poised, 24% of active). Other TFs expressed in mESCs are not highly represented at Cruz Molina enhancers, as it was seen in Whyte enhancers.

In the Pradeepa list, Nanog, Sox2, and Oct4 are abundant at both classes of enhancers (at least one: 46% of canonical, 26% of non-canonical), however often the three TFs are not bound together (~19% canonical, ~7% non-canonical) (Fig 3.7). CTCF is bound preferentially at non-canonical Pradeepa enhancers (~30%) than at canonical ones (~18%), which are in contrast bound by Cohesin (~16% of canonical enhancers, 9% of non-canonical enhancers). Cohesin and CTCF are rarely bound concomitantly at Pradeepa enhancers (~5%). CTCF binding at non-canonical regions was shown in the original Pradeepa *et al*. paper (Pradeepa et al., 2016). As for the other lists, other mESC TFs are not especially enriched at Pradeepa regions.

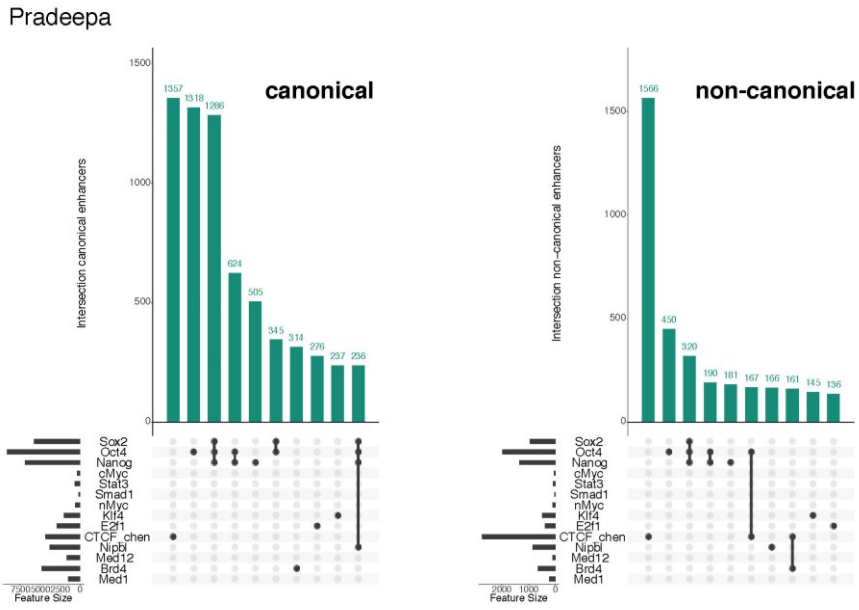


Fig 3..7: Combination of TFs and structural proteins binding at Pradeepa enhancers. a) Plot showing the binding of different proteins at Pradeepa enhancers. On the box below: combination of factors. On the y-axes the number of enhancers bound by the combination indicated. On the x axis on the left: total number of enhancers with that specific protein bound.

When I compare the classes of enhancers across lists, I found that active enhancers (normal, super and active) share binding for Oct4, Sox2, and Nanog, together with an enrichment for Cohesin. Poised enhancers share similarities with active enhancers, such as Nanog, Sox2, and Oct4 binding, whilst primed enhancers show binding mainly for single transcription factors. Pradeepa enhancers, however, show remarkable differences with the other classifications, with a preference for CTCF also at canonical enhancers. Pradeepa enhancers are very long and this could influence the analysis, however this effect is not seen at SEs, which are also longer than Cruz Molina's and normal enhancers.

Taken together, the combinations of protein binding at enhancer lists shows a central role for Nanog, Sox2, and Oct4 at all mESC enhancers considered. These three TFs are preferentially found at active enhancers across lists, while Oct4 alone is mainly found at less active, or repressed, enhancers.

3.4.6 Differently classified enhancers show enrichment for factors used in other lists

The previous analysis of TF binding at different enhancer classes showed similarities between classes of enhancers, together with some striking differences. To understand how chromatin marks and other features relate with classes of enhancers across lists, I also measured the enrichment of other features used as classifiers for the three lists: Nanog, Sox2, Oct4, Med1, H3K27ac, H3K27me3, H3K4me1, H3K122ac, H3K64ac (Fig 3.8).

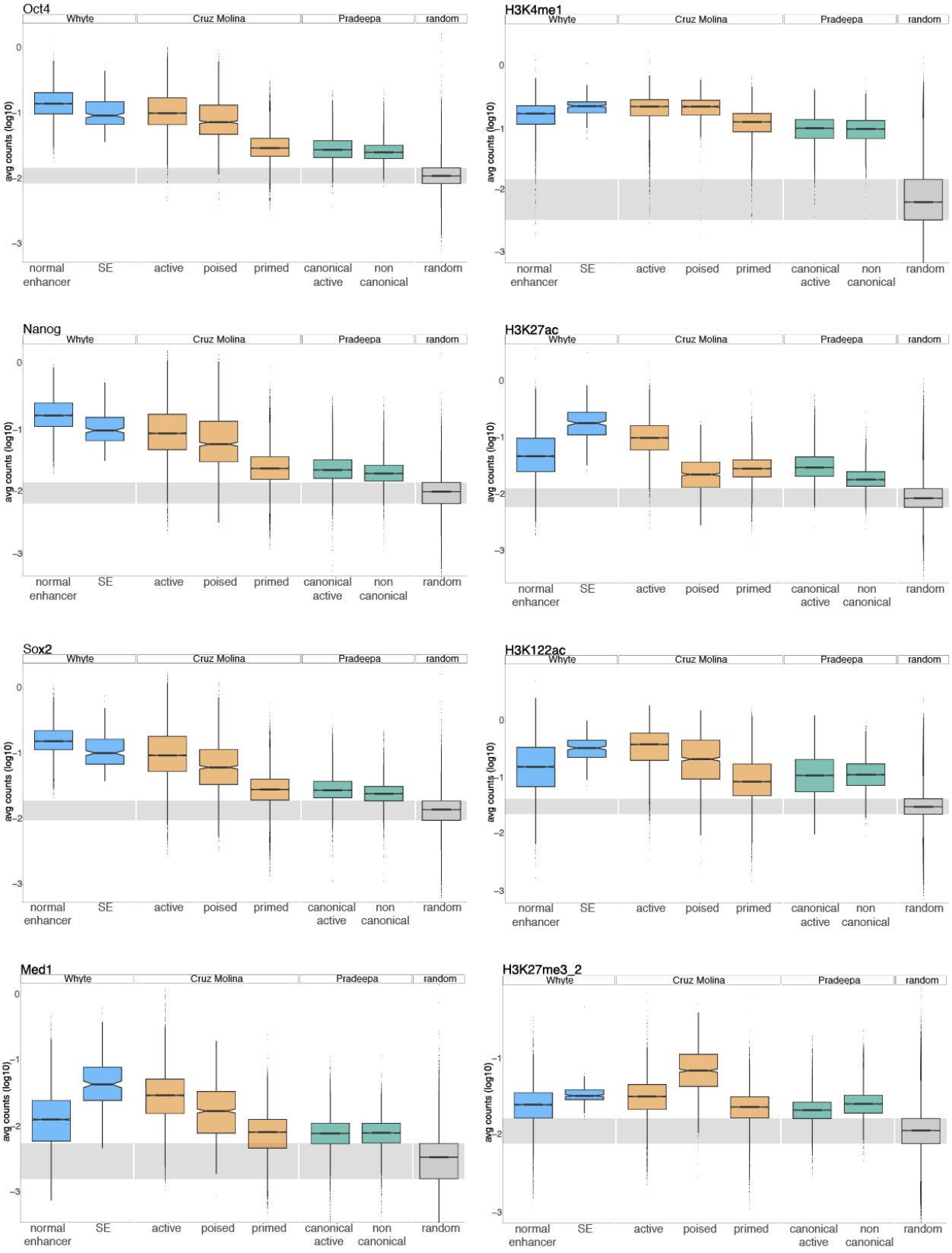


Fig 3.8: Enhancer classes enrichment of classifiers features. a) Boxplots showing the enrichment at enhancer classes for marks used in the enhancer classification. Log of average counts (normalised per length) are represented. A pseudo-count of the minimum count per dataset divided by 10 was added to all counts, to avoid log function of 0. Grey transparent box represents enrichment of the feature at random regions (25%-75% range). Random regions represent randomly shuffled enhancer regions across the genome.

Active enhancers (normal, super, and active) show an enrichment for active enhancer marks: H3K27ac, Nanog, Sox2, Oct4, Med1, and H3K122ac. Notably, SE are the most enriched for H3K27ac and Med1, which are two marks used in the literature to differentiate SE from normal enhancers (Hnisz et al., 2013; Whyte et al., 2013). Poised enhancers are enriched in H3K27me3 and depleted of H3K27ac, while concomitantly being enriched for TF. Primed enhancers show a lower enrichment for all the marks considered, but still above random. H3K4me1 is enriched ubiquitously above random at all classes of enhancers. The Pradeepa lists shows a lower enrichment for all the features compared to the other lists, more similar to the enrichment at primed enhancers for both canonical and non-canonical regions. Canonical enhancers are more enriched for H3K122ac than non-canonical ones; this is in agreement with the original findings in the Pradeepa *et al.* 2016 (Pradeepa et al., 2016). For density distribution of this marks across the two different classes please refer to the Appendix Fig 3.A3. Non-canonical enhancers show a slight enrichment for H3K27me3, which could suggest that a subset of non-canonical enhancers are poised, in line with an observation made in the original paper. The proportion of repressed enhancers among the non-canonical ones was never investigated.

Taken together, the results show that active, SE, and normal enhancers share similar chromatin state together with similar TF binding properties, poised enhancers have a specific chromatin mark, H3K27me3, primed enhancers are in an intermediate or premature state, with less protein binding and feature enrichment that active or poised enhancers. On the other side, Pradeepa's canonical and non-canonical enhancers show mixed characteristics: canonical enhancers show similarities to active enhancers, however with some noticeable differences such as CTCF binding, and non-canonical enhancers show characteristic comparable to primed enhancers.

3.4.7 Enhancer lists identify both unique and shared regions

The analysis performed until now showed that enhancers classified similarly show similar features, however noticeable differences can be found. This prompted me to ask to what extent enhancer regions were identified by multiple approaches. Moreover, it was important to investigate the properties of enhancer regions specific to a given enhancer list. I analysed the enhancer lists co-localization across the genome (Fig 3.9a), considering as co-localisation 1bp of overlap minimum. Many Whyte enhancers (6559/8794, 75%) and Cruz Molina enhancers (21067/32880, 64%) are detected also with other approaches, while many Pradeepa enhancers (16625/32489, 51%) tend to be unique, irrespectively of their longer length. This is in line with the unique properties of Pradeepa enhancers described above, and the shared properties of Whyte and Cruz Molina enhancers, especially for the active class.

To understand whether enhancer regions that co-localise share similar classifications (e.g. normal-active), I analysed the fraction of different classes of enhancers per overlapping group (Fig 3.9b). Indeed, normal, active, canonical enhancers are found to co-localise (in total 1884 overlapping regions out 3391 regions seen in all lists); however, primed and non-canonical enhancers also co-localise with the Whyte list. Poised enhancers, canonical, and non-canonical enhancers are found in different combinations across all the overlapping groups. For the distribution of each class per group of overlap please refer to Appendix Fig 3.A4.

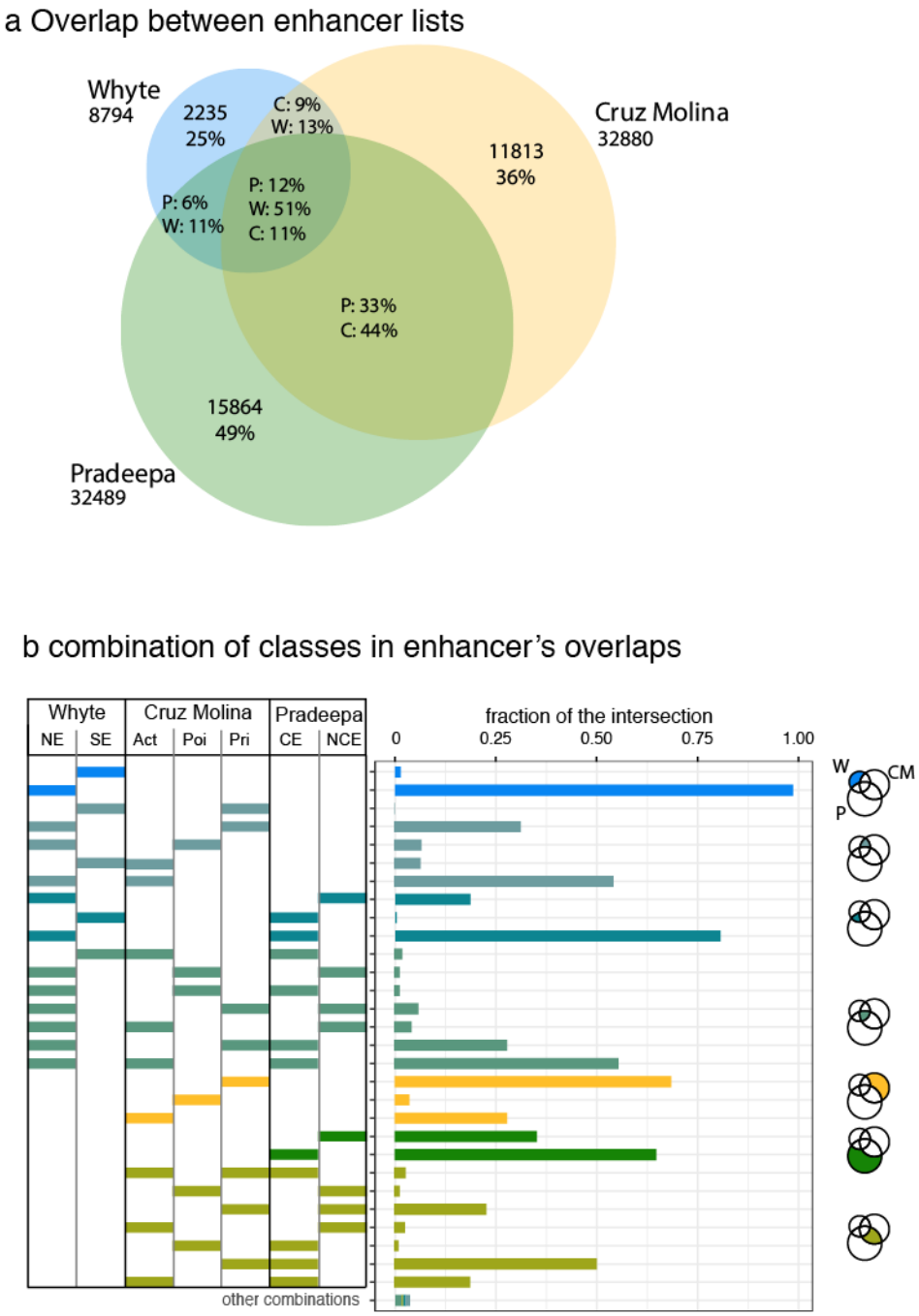


Fig 3.9: Co-localisation of enhancer lists. a) Overlap between Whyte's, Cruz Molina's and Pradeepa's enhancers' location in the genome. An overlap is considered positive if at least 1bp is shared between regions. b) Combination of classes of enhancers per group of overlap. Fraction of group of overlap occupied by each combination of enhancer's classes is shown. On the left, rectangles represent the presence of the specific enhancer class. On the right, schematic of the overlap between different enhancer lists: in colour the group of overlap shown on the graph. W = Whyte, C M = Cruz Molina, P = Pradeepa.

This study shows how complex is the relation between regulatory regions defined with different approaches. Regions classified in different states co-localise in the genome, making difficult to estimate which approach or classification defines and classify enhancers more reliably.

3.4.8 Candidate enhancer regions identified by different criteria have heterogeneous features

Enhancers identified in different lists co-localise in the genome. One possibility would be that overlapping enhancer regions could share chromatin features, while uniquely found regions could have specific ones. I therefore investigated the different groups of enhancer regions according to whether they were unique or common to different enhancer lists, and explored their occupancy features. The overlapping enhancer regions were merged and compared for their relative normalised ChIP-seq enrichment of different marks (Fig 3.10a). The absolute ChIP-seq enrichment for selected features at all overlapping regions can be found in Appendix Fig 3.A5 .

Figure 3.10

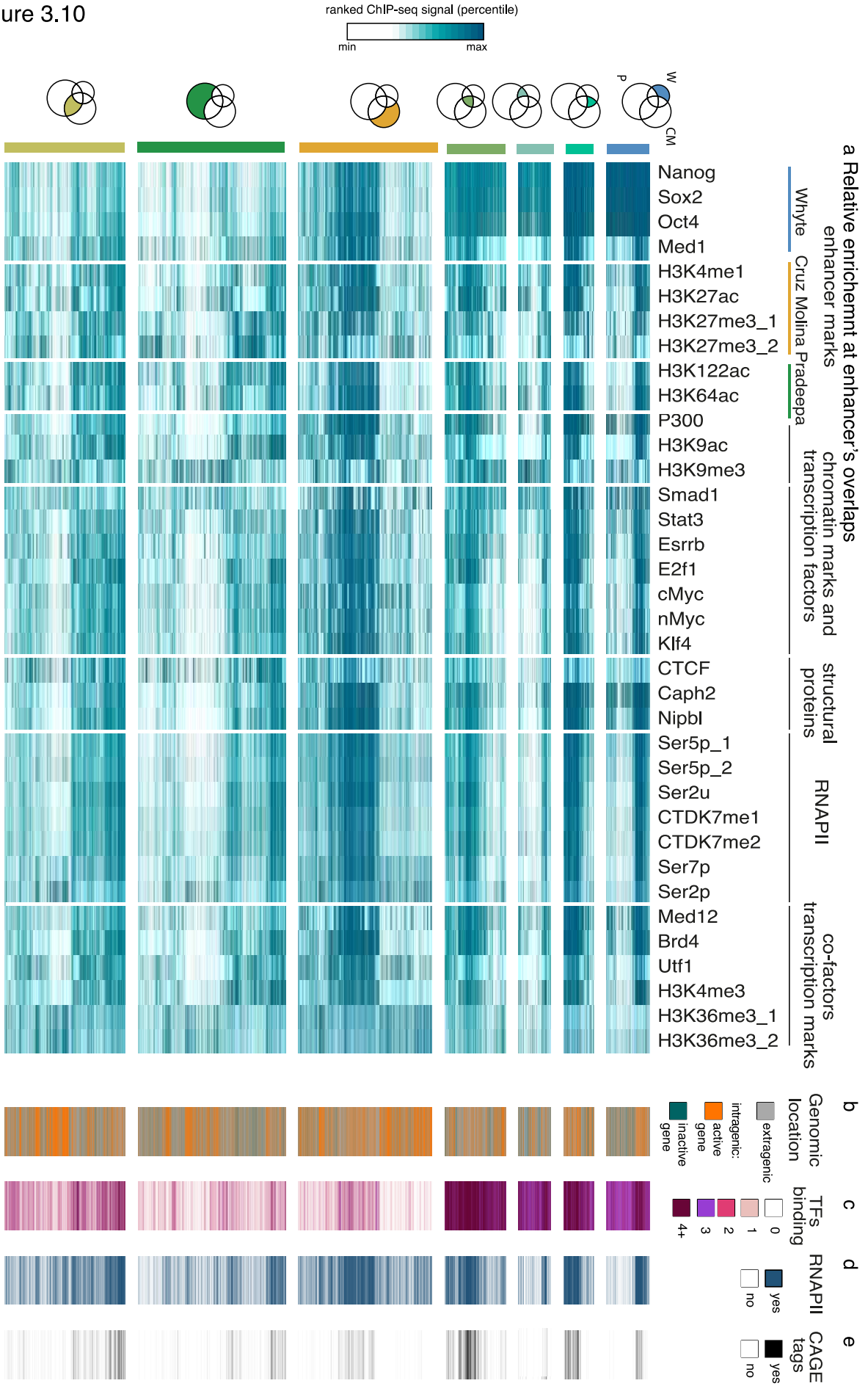

ranked ChIP-seq signal (percentile)
min
max
W
P
CM
a Relative enrichemnt at enhancer's overlaps
enhancer marks
Whyte
Cruz Molina
Pradeepa
chromatin marks and transcription factors
structural proteins
RNAPII
co-factors transcription marks
Nanog
Sox2
Oct4
Med1
H3K4me1
H3K27ac
H3K27me3_1
H3K27me3_2
H3K122ac
H3K64ac
P300
H3K9ac
H3K9me3
Smad1
Stat3
Esrrb
E2f1
cMyc
nMyc
Klf4
CTCF
Caph2
Nipbl
Ser5p_1
Ser5p_2
Ser2u
CTDK7me1
CTDK7me2
Ser7p
Ser2p
Med12
Brd4
Utf1
H3K4me3
H3K36me3_1
H3K36me3_2
b Genomic location
extragenic
intragenic: active gene
inactive gene
c TFs binding
0
1
2
3
4+
d RNAPII
no
yes
e CAGE tags
no
yes

Fig 3.10 : Overlapping enhancer regions show diverse features enrichment. a) Percentile of enrichment of ChiP-seq normalised (ranked) reads of different group of features at overlapping groups of enhancers. Overlapping groups are schematised on the left of the graph. W = Whyte, C M = Cruz Molina, P = Pradeepa. Datasets are divided in broad categorised specified above the graph. Darker colour means stronger relative enrichment, lighter colour means lower relative enrichment in all the overlapping regions. b) Heatmap showing the genomic location of overlapping regions. Extragenic regions: outside +/- 2kb from TSS-TES of annotated genes, in orange; intragenic active gene: +/- 2kb from TSS-TES of annotated active genes (TPM >1), in blue; intragenic inactive gene: +/- 2kb from TSS-TES of annotated inactive genes (TPM <1), in white. TPM calculation where taken from published tables on mESCs from Ferrai et al. 2017 17. c) Heatmap showing the number of TF peaks at overlapping enhancer regions. TF binding where binned in: 0 TF peaks (white), 1 TF peak (light pink), 2 TF peaks (pink), 3 TF peaks (violet), >=4 TF peaks (dark purple). d) Heatmap showing the presence of RNAPII peaks peaks at overlapping enhancer regions. In dark blue: overlap with a RNAPII peak, in white, no overlap with a RNAPII peak. e) Heatmap showing the presence of CAGE tags outside promoters at overlapping enhancer regions. In black: overlap with a CAGE tags, in white, no overlap with a CAGE tags. CAGE-tags were downloaded from the FANTOM5 website) http://fantom.gsc.riken.jp/) by Markus Schueler. Every row in all the heatmaps represents an overlapping enhancer region.

All subgroups of overlapping or unique enhancer regions show heterogeneous enrichment for classification features. For example, regions common to all three enhancer lists have a high enrichment for Nanog, Sox2, and Oct4, but show variable H3K27ac, while others contain H3K27me3. In general, regions overlapping Whyte enhancers have the strongest normalised ChIP-seq enrichment signal for Nanog, Sox2 and Oct4, but similar strength of enrichment for the same three TFs can be seen in the Cruz Molina-only regions. Regions uniquely identified by each of the three approaches also show heterogeneity: Pradeepa unique regions have low enrichment for Nanog, Sox2 and Oct4, can be high or low in H3K122ac, and are enriched in CTCF, which was expected from the original Pradeepa analyses.

Another example of the heterogeneity is for example in the regions shared by Cruz Molina and Pradeepa, which show in general lower enrichment for all the marks considered, but can be associated with features of active genic regions, such as H3K36me3 and RNAPII-S2p. To understand whether the presence of this latter signature results from active transcription at coding regions, I took advantage of published gene expression levels in mESC obtained by RNA-seq (Ferrai et alFerrai et al., 2017) and categorized all the regions as: extragenic, if outside +/- 2kb annotated TSSs and TESs; intragenic active, if inside +/- 2kb annotated TSSs and TESs of an active gene (TPM >=1); intragenic inactive, if inside +/- 2kb annotated TSSs and TESs of an inactive gene (TPM <1). Indeed, active transcribing chromatin mark H3K36me3 appears in regions inside active genes (Fig 3.10b). Raw read enrichment of RNAPII is visible at extragenic regions (Fig 3.11).

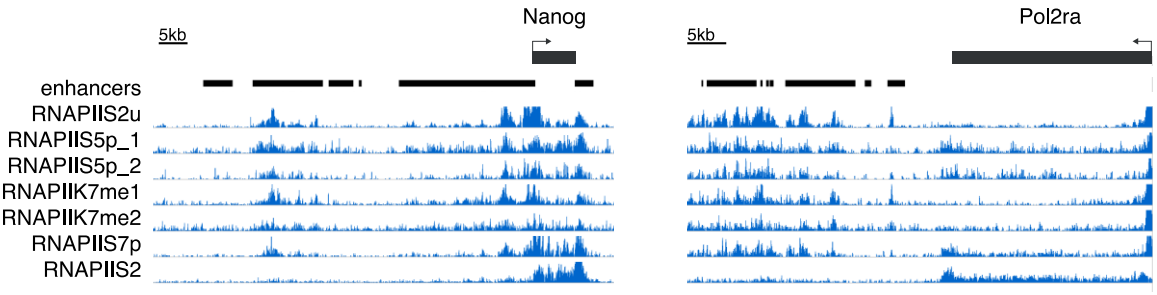


Fig 3.11: Single gene examples of RNAPII coverage. Images generated with IGV software.

The picture that emerges from this enrichment analysis is extremely complex. Therefore, to investigate whether TF binding could help prioritise specific enhancer groups, I measured the number of TF found to occupy each enhancer region and whether this binding could be related with specific chromatin features, I counted the number of TF bound at least once per region. I first calculated the distribution of binding events inside the whole dataset of overlapping regions to define optimal binning (Appendix Fig 3.A6); regions were then categorized into 0, 1, 2, 3 or ≥4 TFs bound (Fig. 3.10c). Unexpectedly, a great number of common enhancer regions are not bound by any of the TFs studied (25558/46945, 54%). Candidate enhancer regions inside active genes have fewer TF binding events than extragenic regions (~50% in extragenic regions, ~40% in active gene regions, 46% in inactive gene regions).

3.4.9 RNAPII occupies regions bound by one or more transcription factor

The association of RNAPII at enhancers has previously been shown (Cruz-Molina et al., 2017; De Santa et al., 2010; Kaikkonen et al., 2013). To explore whether RNAPII binding could help characterize the different enhancer regions, especially those occupied by TFs, I quantified the presence of RNAPII and of CAGE tags at these regions (Fig 3.10d-e). CAGE tags were previously used to identify enhancers, as an orthogonal approach to chromatin marks based on eRNA transcription and mark active transcription outside promoters (Arner et al., 2015). RNAPII is present at 50% of all enhancer regions considered here, of which 42% are extragenic. The regions where at least 1 TF are present have 50% change of being occupied by RNAPII, whereas most (87%) of regions occupied by more than 4 TF have RNAPII, even when only extragenic regions are considered (Fig 3.12a). The CAGE tags dataset is not numerous (2604), however CAGE tags are preferentially found in regions with >3 TFs (1283/2604, ~50%) (Fig 3.10e), and at these region RNAPII is bound 98% of the time. Therefore, when numerous TFs are present at a regulatory region, RNAPII is also bound. The transcriptional activity of

extragenic RNAPII occupancy at candidate enhancer regions will be investigated in more detail later in this thesis.

To assess whether the presence of RNAPII might relate with TF abundance, I took all the extragenic overlapping enhancer regions (outside +/- 2kb from annotated TSSs and TESs) with one TF bound, divided the read distribution in quartiles and quantified the fraction of region bound by RNAPII per quartile (Fig3.12b, 6 TFs analysed are shown). A trend is visible where the more a TF is enriched (higher quartile) the more RNAPII occupies that region (higher fraction). Interestingly, Smad1 shows a different behaviour of anti-correlation between the quantity of TF bound and the presence of RNAPII, which could suggest a repressive activity of Smad1 at enhancers. Other Smad proteins, such as Smad4, where shown to have a repressive function (Vincent et al., 2009; Wotton et al., 1999).

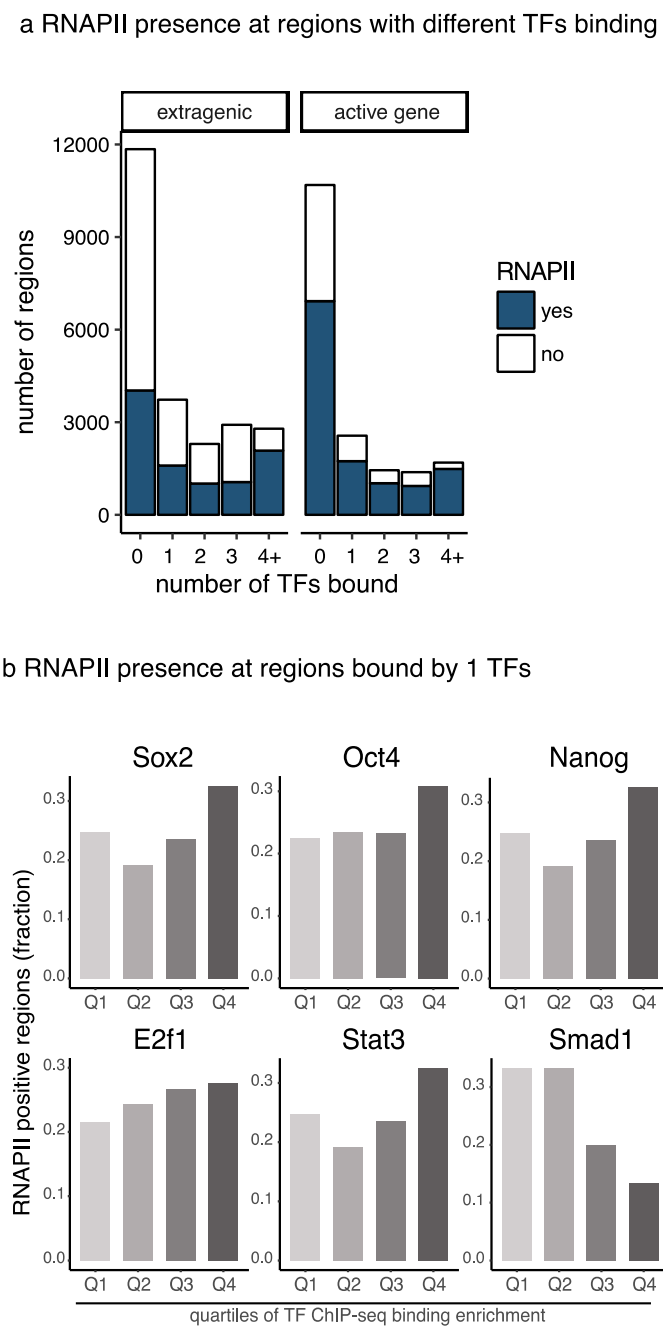


Fig 3.12: RNAPII binding at overlapping enhancer regions co-occurs with high TF binding. a) RNAPII binds at overlapping regulatory regions when >=1 TF is present. In blue: regions overlapping with a RNAPII peak; in white, regions not overlapping with a RNAPII peak. Regions are divided for their genomic location: extragenic region are +/-2kb from TSS-TES of annotated genes; active genes inside SS-TES of annotated active genes (>1 TPM). TPM calculation where taken from published tables on mESCs from Ferrai et al. 2017. b) RNAPII preferentially binds at overlapping enhancer regions with high TF enrichment. Barplot showing the percentage of RNAPII positive enhancer overlapping regions per quantile of enrichment of Sox2, Oct4, Nanog, E2f1, Stat3, Smad1. Only extragenic regions are shown: +/-2kb from TSS-TES of annotated genes.

3.4.10 RNAPII is present in different activation states at regulatory regions

Post-translational modifications of the CTD of RNAPII are informative of its state of activation. To explore whether RNAPII is present at enhancers in different activation states and whether the activation correlates with enhancer features, namely TFs, I analysed the co-occupancy of different RNAPII modifications and TFs at regions characterized by binding of 1, 2, 3 or 4 TFs. To avoid confounding effects, only extragenic regions were considered in this analysis (outside +/- 2kb from annotated TSSs and TESs). All the different forms of RNAPII considered in this analysis were found present at regulatory regions to different extents, including RNAPIIS5p (characteristic of active and poised gene promoters), RNAPIIS7p (of active promoters) and RNAPII-S2p (of elongation; Fig 3.13a). The extent of occupancy of RNAPII modifications at extragenic enhancer regions reflects the RNAPII activation state, with less active RNAPII found

more often than fully elongating one at these regions (~46% RNAPIIS2u (8WG16ab), ~37% RNAPIIS5p, ~30% RNAPIIS7p, ~25% RNAPIIS2p).

Figure 3.13

Most represented combinations of TFs and RNAPII modification at extragenic overlapping regions divided by number of TFs bound

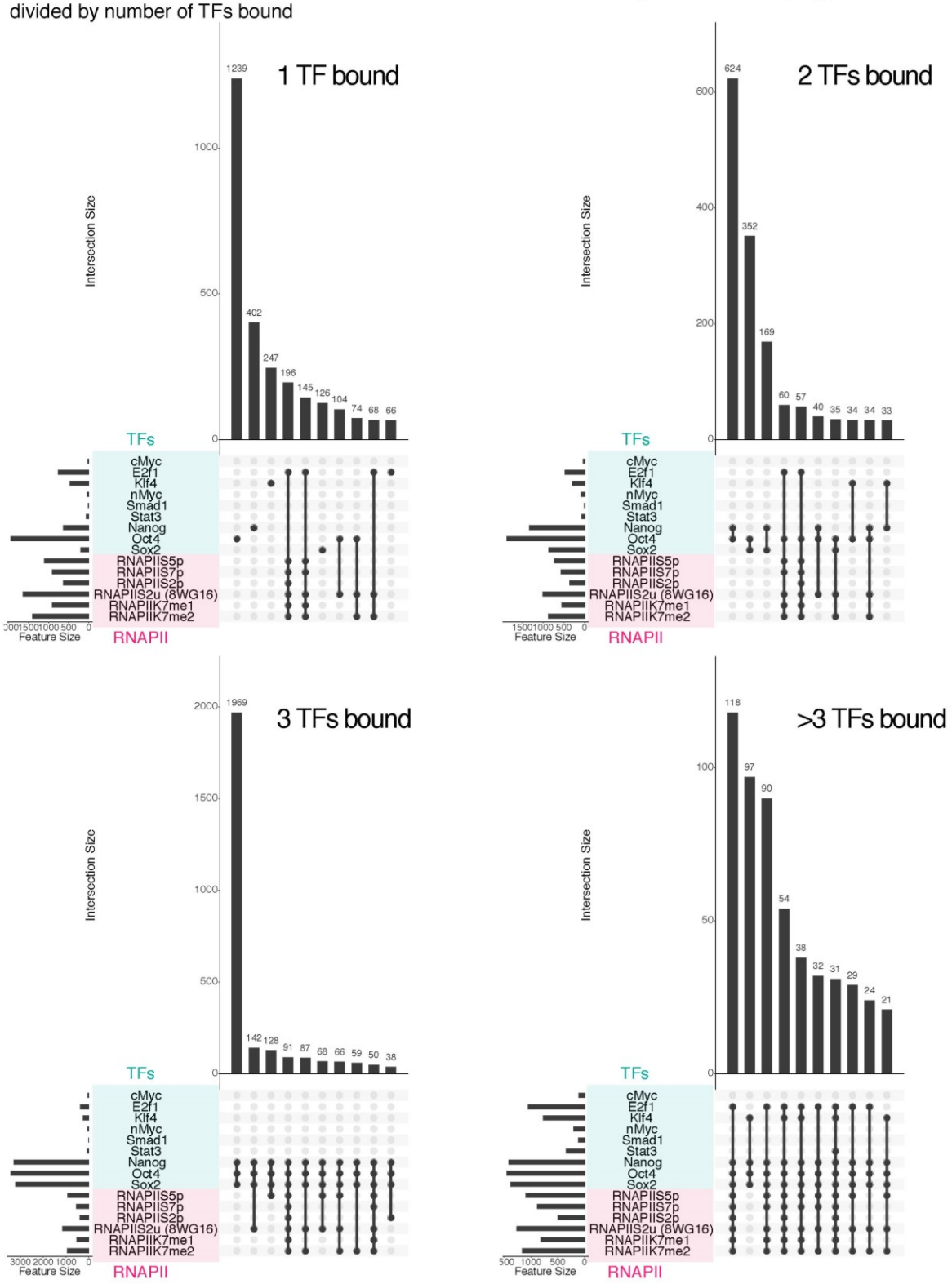


Fig 3.13: Most represented combination of TFs binding and RNAPII modifications at extragenic overlapping enhancers. a) Combinations of TFs and RNAPII modifications peaks at extragenic overlapping enhancers, showed by number of TFs bound per region: 0 TF peaks, 1 TF peak, 2 TF peaks, 3 TF peaks, >= 4 TF peaks. Only extragenic regions are shown: +/-2kb from TSS-TES of annotated genes. Blue transparent box around features represents TF datasets; pink transparent box around the analysed features represents RNAPII datasets.

To understand whether the extent of TF binding positively correlated with RNAPII recruitment and activation state, I correlated the ChIP-seq reads of different RNAPII modifications and TFs at extragenic overlapping regulatory regions (Fig 3.14). RNAPII modifications are highly correlated with themselves, as expected (e.g. Dias *et al.* 2015 (Dias et al., 2015)), with E2f1 and partially with Klf4. Nanog, Sox2 and Oct4 mostly correlate with themselves. This result, connected with the finding that 1969 regions show a unique biding of Nanog, Sox2 and Oct4 without any RNAPII (Fig 3.12a), suggests that RNAPII regulation is not directly linked to Oct4, Nanog and Sox2 binding, but could be recruited after TF binding specifically at some enhancer regions, to participate in specific aspects of enhancer activity. Interestingly, similar correlation was found at promoters (+/- 2kb from TSSs) of active genes (Fig 3.14b) suggesting that promoters and enhancers share common features. Interestingly, the correlation between Sox2, Nanog and Oct4 at TSS is lower than the one at enhancers. Correlation over all promoter regions can be found in Appendix Fig 3.A7.

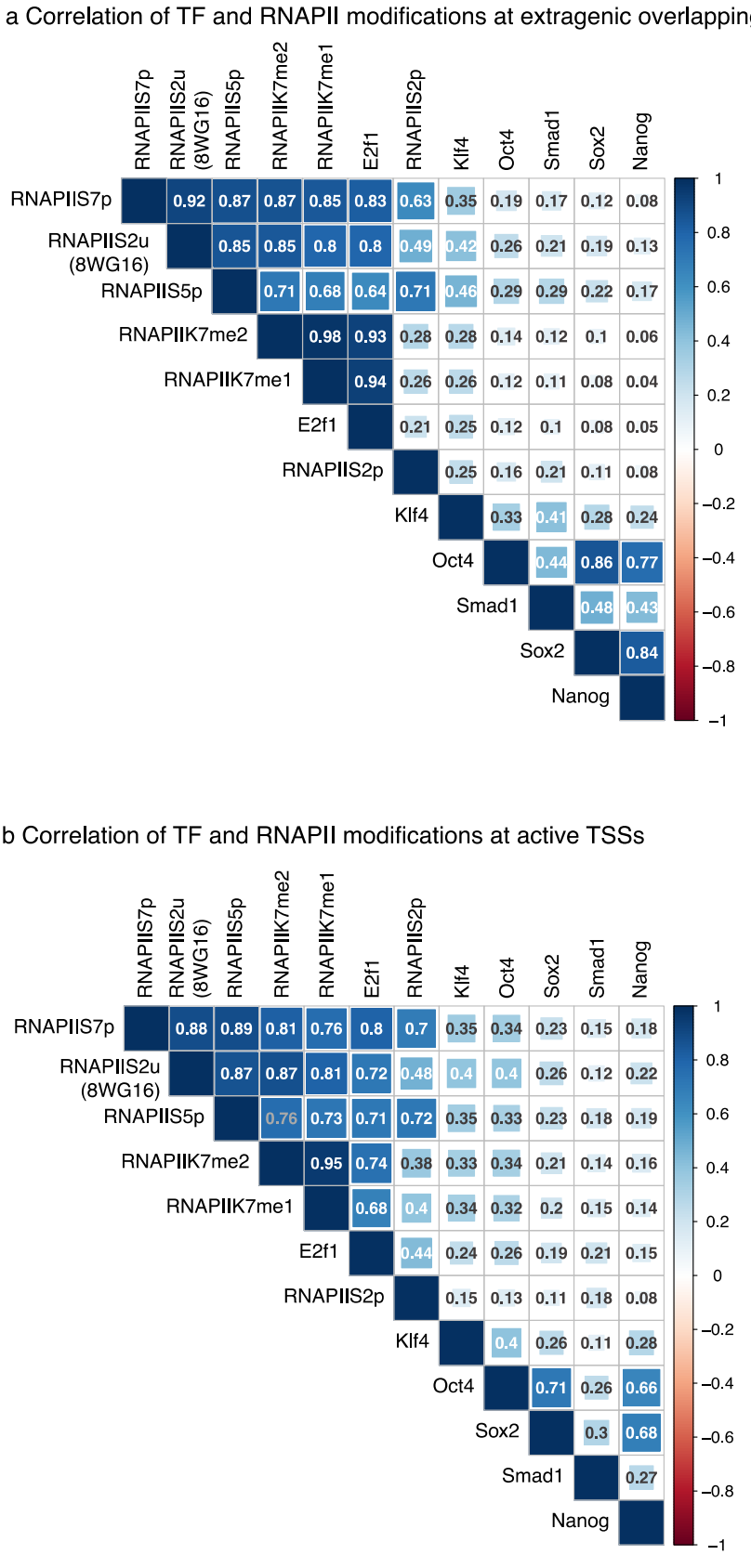


Fig 3.14: RNAPII modifications correlate with TFs at extragenic overlapping enhancers a) at extragenic enhancers and b) at active TSSs. Correlation plot showing pearson correlation between RNAPII modifications and TFs. Plot made with the corrplot package in R, and hierarchical clustering option.

3.5 Discussion

After their first discovery (Banerji et al., 1981), enhancers were extensively investigated and described in different states of activation. These states can be defined with different combinations of chromatin occupancy features and are thought to describe the level of activity and diverse functions of enhancers (Calo and Wysocka, 2013). In parallel, studies have clarified the important role of transcription factors in target gene activation (Spitz and Furlong, 2012), and the role of proteins such as CTCF and Cohesin in mediating chromatin loops; loops create a preferred region of interaction where enhancers might act (Rao et al., 2014). However, the enhancer field lacks a clear comparison between the different classes and states of activation of enhancers. Some studies have tackled these questions both regarding methodologies of enhancer identification (Kleftogiannis et al., 2014) and basic enhancer features (Benton et al., 2017), however they focused only on active enhancers leaving out the comparisons of different classes of enhancers defined with different approaches or associated with poised states.

In this chapter, I show that enhancer classes identified by different approaches are found at distinct genomic locations. Regions more enriched for typical enhancer marks, therefore more likely to act as enhancers *in vivo*, are more enriched in TFs. When one or more TF is bound to a regulatory region, 50-90% of the time RNAPII would be bound. RNAPII is therefore an interesting candidate to investigate regulatory regions.

3.5.1 Enhancers lists identify regions with different features

To understand how different enhancer classification approaches relate with each other, I compared lists and classes of enhancers in mESCs from three publication: Whyte *et al. (Whyte et al., 2013)*, Cruz Molina *et al. (Cruz-Molina et al., 2017)*, Pradeepa *et al. (Pradeepa et al., 2016)*. The three lists I analysed classify enhancers in five different states: active/normal/canonical, super, poised, primed and non-canonical. These different classes were described in the literature to have distinct functions and characteristics: normal (Whyte list)/active (Cruz Molina list)/canonical (Pradeepa list) enhancers enhance target gene expression, possibly through chromatin looping, and are marked by H3K4me1 and H3K27ac, or bound by specific TFs, such as Nanog, Sox2 and Oct4 (Calo and Wysocka, 2013; Zentner et al., 2011). Super enhancers (Whyte list) are clusters of enhancers or stretched enhancers that can act synergistically; they cover a large distance in the genome and greatly enhance target gene expression. Super enhancers are marked by high levels of Med1 or H3K27ac and bound by specific TFs (Hnisz et al., 2013; Whyte et al., 2013). Poised enhancers (Cruz Molina list), on the other hand, repress their target genes in mESC via direct chromatin contact and activate them during differentiation,

when they lose the repressive mark H3K27me3 and acquire the active mark H3K27ac (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). Primed enhancer (Cruz Molina list) are not yet active and marked by H3K4me1 and pioneer transcription factors (Calo and Wysocka, 2013). Finally, non-canonical enhancers (Pradeepa list) were recently found to enhance target gene expression similarly to active enhancers, without being marked by the canonical enhancer mark H3K27ac and instead being marked by H3K122ac. In the original paper was observed that a subgroup of non-canonical enhancers is marked by H3K27me3, which is coherent with some of the results in the current chapter (Fig 3.8).

These different enhancers are widespread through the genome, however their distribution at gene loci shows discrepancy between the classifications: very active enhancers, such as super enhancers, could be located at the same regions were primed enhancers, or even poised enhancers are located. Partially, this could be due to the size differences between enhancer classifications: super enhancers tend to be longer, as close TF peaks used to defined enhancers were merged in larger regions in the original paper (Whyte et al., 2013). Cruz Molina's regions on the other hand are the shortest, while Pradeepa's enhancers are intermediately long. The fact that SE can be found to contain enhancers in different states could suggest that these regions contains numerous enhancers that potentially don't share the same activation state and could have different functions or respond to different stimuli which should act in concert, possibly in a big chromatin domain together with co-regulated genes.

Different classes of enhancers show diverse TFs binding and histone modification enrichment: active/normal/super enhancers have simultaneous binding of Nanog, Sox2 and Oct4. These three TFs are frequently bound to the enhancers examined in this study and represent pivotal TFs in mESC that regulate cell stemness (Kashyap et al., 2009). Interestingly, less active classes of enhancers are bound more often by Oct4 alone. Oct4 was shown to be a pioneer TF (King and Klose, 2017), in accordance with his presence at primed enhancers. Preadeepa enhancers are the least enriched for enhancer marks across all the enhancers considered, also when the non-canonical mark H3K122ac is analysed. Moreover, Pradeepa's regions show a preference for CTCF, which is shared to some degree with primed enhancer, but not found at active enhancers. In particular, canonical enhancers were defined as active and, even if they were classified with the same logic of active enhancers in the Cruz Molina list, substantially differ from the other active enhancers also regarding length. These differences can be attributed to technical reasons, such as threshold choice, or peak calling softwares used. However, this observation raises concerns on the robustness of enhancer identification approaches and their reproducibility.

Differences in features between enhancer classes could be indicative of diverse functions; regulatory regions might act through different mechanisms, while still regulating target gene expression. Specific mechanisms were shown for poised enhancers, that repress their target genes (Cruz-Molina et al., 2017), or for active genes contacting the target gene via chromatin looping(Ferrai and Pombo, 2009) or via RNAPII tracking from the enhancer to the promoter (Wang et al., 2005). Some of the regulatory regions investigated in this chapter could be mainly involved in looping of chromatin and therefore show a higher occupancy for proteins such Cohesin, while other regions could act as dock for transcription factors and co-factors or might be involved in chromatin structure, if for example bound by CTCF. The different scenarios can co-exist, where regulatory regions in close proximity to a gene create a highly dense TFs hub and a preferred region of interaction.

3.5.2 Enhancer lists co-localisation is complex

To understand the relationship between the different enhancer lists I look at their co-localisation and analysed the resulting overlapping groups. The enhancer regions identified in the 3 different publications are not completely recovered in any other list. Every approach identifies unique regions, which are however not only the unique class of that particular list. For example, the majority of Whyte enhancers are found also in either Cruz Molina or Pradeepa lists, however the regions unique to the Whyte lit are not only super enhancers. This may be surprising, as super enhancers can be categorised as active enhancers or canonical enhancers in the other two lists. The regions unique to Cruz Molina are not solely poised and primed, and surprisingly poised enhancers where also found in different lists that where not considering this repressed state. The internal differences in groups can be attributed to different classes of enhancers that co-localise between lists and are therefore assigned to the same group. All lists considered in this study have classes not considered in other studies, which makes it all more surprising that one region classified as active with one approach could be classified as primed or poised with another.

Surprisingly, numerous enhancers showed no TFs binding, which is known to be important for enhancer activation. It is possible that these regions are bound by other TFs not considered in this study. However, another intriguing possibility is that these regions show a diverse function, may be bound by structural protein such as CTCF and contribute to genome architecture and are less involved in direct gene regulation.

The results highlight how heterogeneous the landscape of enhancers in mESC is. Different approaches not only define different regions with different characteristics, but also classify them in different states of activation. The heterogeneity is preserved even if co-localising regions are

grouped; groups have diverse characteristic and features within themselves. Interestingly, regions identified by more than one approach do not appear in general terms more enriched for enhancer marks, compared with regions unique to one study. This is in accordance with the study from Benton and colleagues (Benton et al., 2017) where active enhancers identified with different approaches where shown to be not more active than enhancers identified by single approaches.

Enhancer identification and classification remains a challenge in the field; further studies should aim to clarify what kind of regions every approach identifies, what is the power of the approaches and how different types of enhancers act and are regulated.

3.5.3 Transcription Factor binding co-occurs with different forms of RNAPII

RNAPII transcription at enhancers was indirectly shown to be important for their activity: eRNAs produced at enhancers correlate with mRNA produced at target genes (Kim et al., 2010; Kaikkonen et al., 2013). To investigate RNAPII at enhancers I analysed its binding in relation with TFs. Enhancers were showed to act as recruitment sites for TFs (Spitz and Furlong, 2012), increasing their abundance near target genes in the 3d space of the nucleus; a similar mechanism was suggested for RNAPII creating a hub of co-regulated genes. Although only a limited number of TFs were analysed in the current chapter, RNAPII binds at regulatory regions together with transcription factors in at least 50% of the regions considered. RNAPII binding is influenced by the abundance of TF at the regulatory region: more TF occupies a regulatory region more likely it is that RNAPII would be bound as well. RNAPII binds at regulatory regions when different TFs are present in different forms, here analysed via its post-translational modifications of its CTD. TFs could bind at regulatory regions and act as recruiter for RNAPII, similarly to what happens at genes (Brookes and Pombo, 2009). RNAPII recruited at enhancers could be post-transnationally modified and transcribe eRNAs. Active RNAPII states are found at regulatory regions, especially concomitantly with 4 or more TFs bound, which suggest a link between RNAPII state and enhancer state. TFs can bind in combinations, which in turn can influence the state of the regulatory region (Merika and Thanos, 2001). This phenomena, added to the observation that RNAPII can be found in different states at enhancers, amplify the regulatory landscape of enhancers: combinations of TFs and RNAPII states can finely tune enhancer activity and consequently target gene expression.

However, RNAPII doesn't bind all regions bound by a TF. Smad1 has little to no co-binding together with RNAPII at regulatory regions. RNAPII and Smad1 could be mutually exclusive

and potentially Smad1 could negatively regulate enhancer regions. It would be interesting to further understand the relation between this TF and RNAPII at regulatory regions and to understand if other TFs behave in the same way and what are their characteristics, how is this repressive process regulated, and what are the target genes.

Interestingly, a high correlation between RNAPII and E2f1 was found. E2f1 is a master regulatory transcription factor involved in cell cycle and up regulated in cancer. Its targets were shown to be down regulated after Brd4 inhibition that affected specifically SE in cancer cells (Chapuy et al., 2013), and it was suggested that it is involved indirectly in RNAPII-CTD phosphorylation (Ma et al., 2003). A more detailed exploration of the relation between E2f1 and RNAPII at extragenic candidate enhancers would be of great interest however was beyond the aims of this thesis.

The fact that RNAPII is present at enhancer, possibly in different states of activation, could suggests that genes and enhancers share RNAPII and TFs, possibly in specialised chromatin hubs. The TFs combination and the RNAPII states can be regulated concomitantly in the hub, leading to a coordinate gene regulation. However, no study dealt yet with the activation state of RNAPII at enhancers and the implication on enhancer state and their activity on gene regulation. The study of RNAPII regulation could be beneficial for understanding the state of enhancers, clarify their mechanism of activation and identify enhancers in the genome.

In the next chapters, I investigate if RNAPII is present in different states at enhancers, if these states resemble the ones at genes, and if extragenic RNAPII can be used as a tool to find enhancers and their state.

3.6 Figures Appendix

Examples of gene regions with differently define enhancer classes

Fig 3. A1: Examples of gene loci regulatory landscape. a) Enhancer classes locations at 3 gene loci. Whyte, Cruz Molina and Pradeepa enhancers and RefSeq genes are shown. Classes of enhancers are indicated below the enhancers and color-coded: Whyte enhancers: normal – light blue, super – dark blue; Cruz Molina enhancers: active – brown, poised – orange, primed – gold; Pradeepa enhancers: canonical: light green; non-canonical – dark green. Images generated via the IGV software.

Figure 3.A2

Combination of TFs at different classes of enhancers

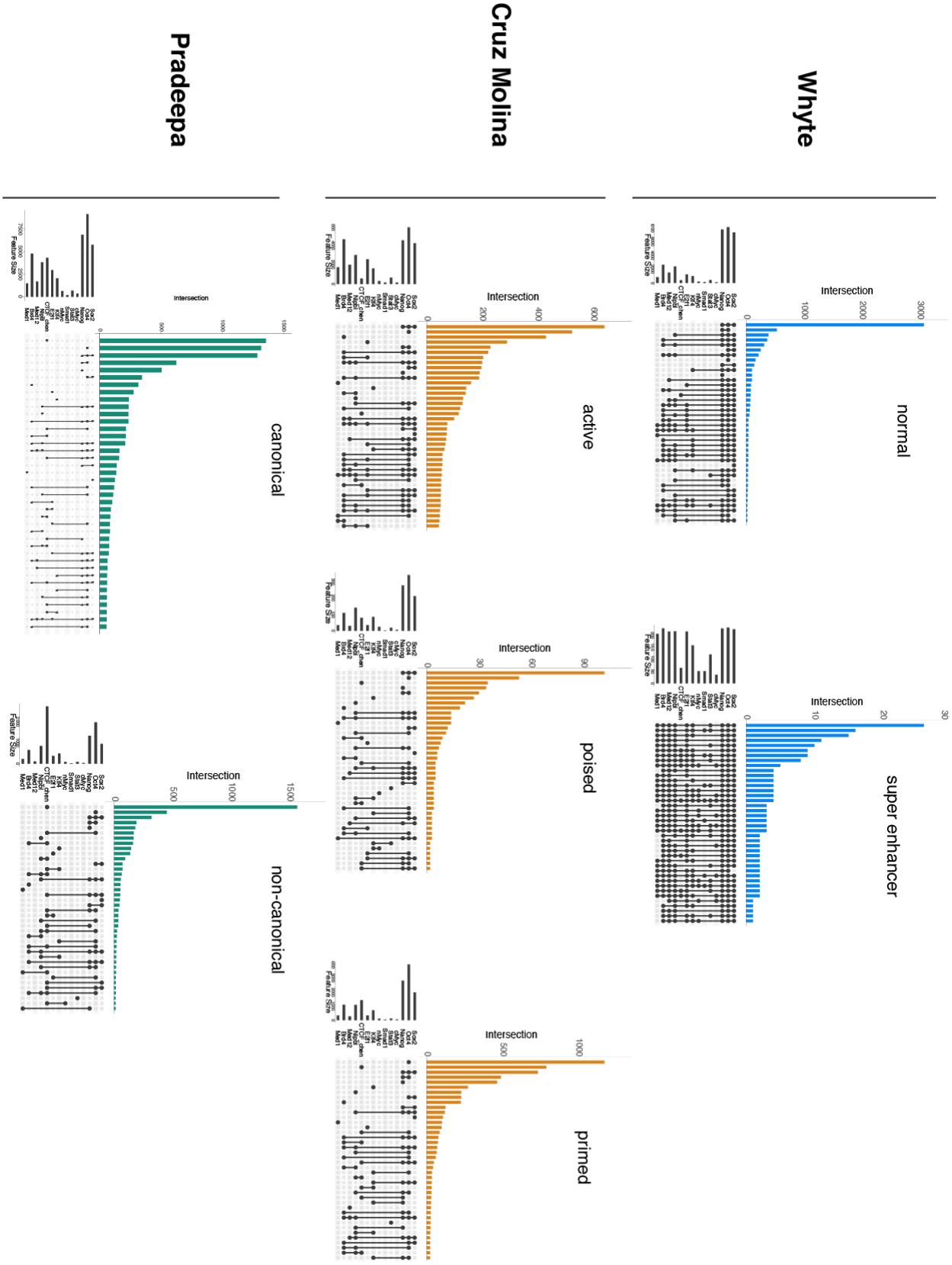


Fig 3.A2: All combinations of different TFs and structural protein at enhancers. Plot showing the binding of different proteins at enhancer lists. On the box below: combination of factors. On the y-axes the number of enhancers bound by the combination indicated. On the x axis on the left: total number of enhancers with that specific protein bound.

Density of enrichment of H3K122ac and H3K64ac at canonical and non-canonical enhancers from Pradeepa et al

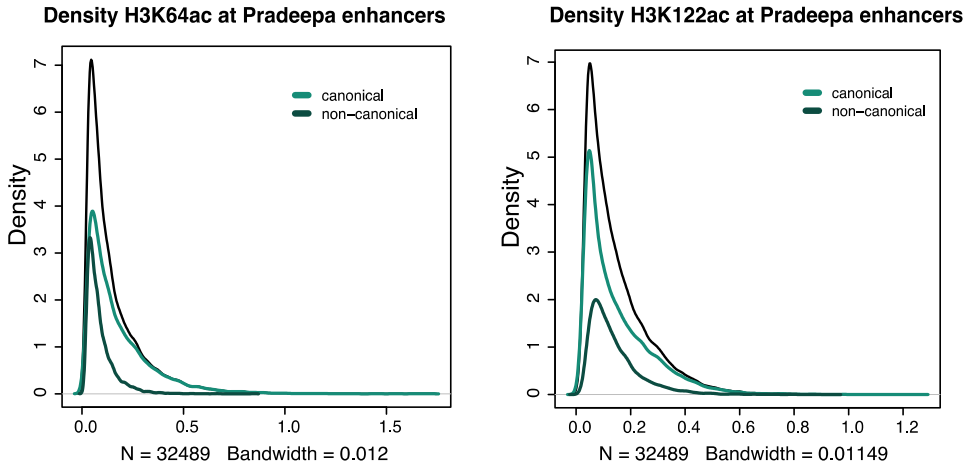


Fig 3.A3: Density of enrichment of non-canonical marks at Pradeepa enhancers. Black line: density of enrichment at all Pradeepa enhancers; dark green: density of enrichment at non-canonical enhancers; light Green: density of enrichment at canonical enhancers. Density calculated with the density function in R.

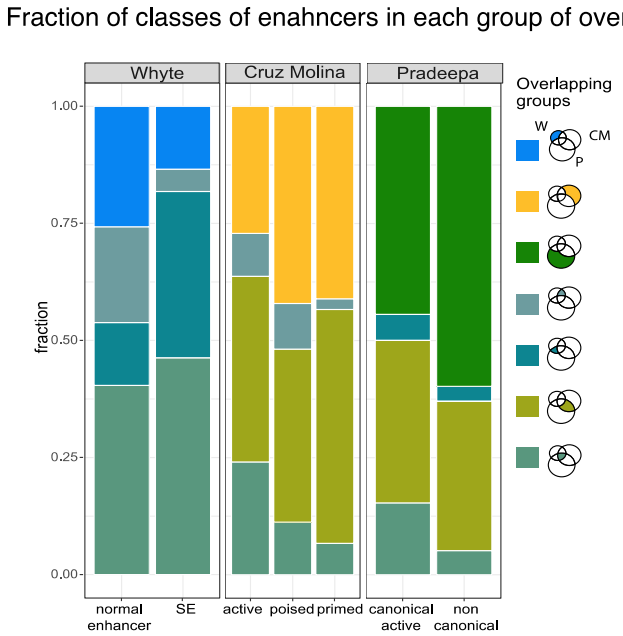


Fig 3.A4: Distribution of enhancer classes per group of overlap. a) Plot showing the fraction of different classes of enhancer present in the overlapping group. Legend of colours representing groups is on the right of the plot, together with a schematic representing the overlapping group. W = Whyte enhancers, CM = Cruz Molina enhancers, P = Pradeepa enhancers.

Figure 3.A5

Enrichment of different factors at overlapping regions

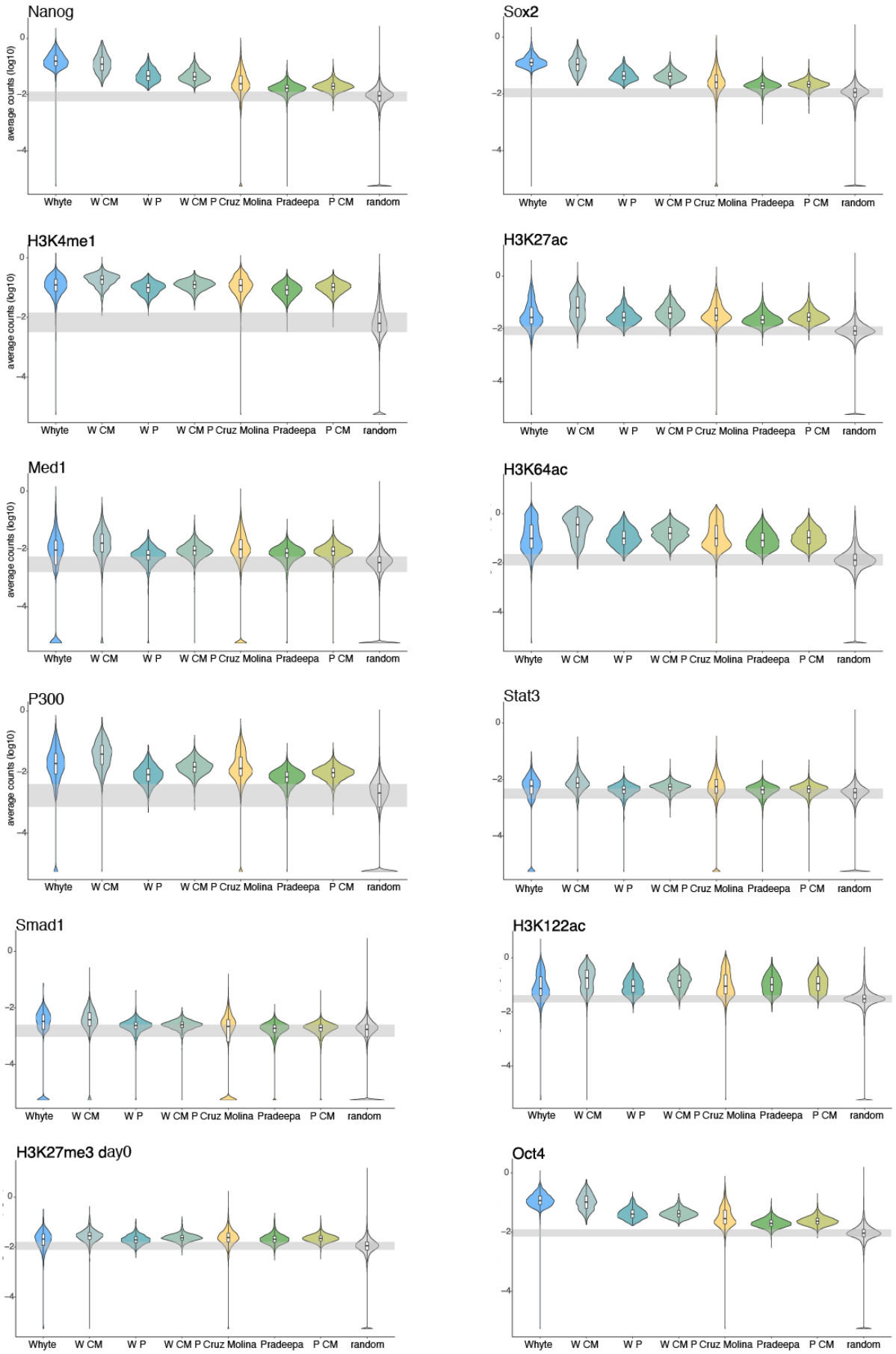


Fig 3.A5: Enrichment of different features at overlapping groups of enhancers. a) Violin plots showing the distribution of absolute enrichment of feature per overlapping groups. Log of average counts (normalised per length) are represented. A pseudo-count of the minimum count per dataset divided by 10 was added to all counts, to avoid log function of 0. Grey transparent box represents enrichment of the feature at random regions (25%-75% range). Whyte = Whyte enhancers unique regions; W CM = Whyte - Cruz Molina overlapping regions; W P = Whyte -Pradeepa overlapping regions; W CM P = Whyte - Cruz Molina - Pradeepa overlapping regions; Cruz Molina = Cruz Molina unique regions; Pradeepa = Pradeepa unique regions; P CM = Pradeepa – Cruz Molina overlapping regions. Random Regions = randomly shuffled overlapping regions.

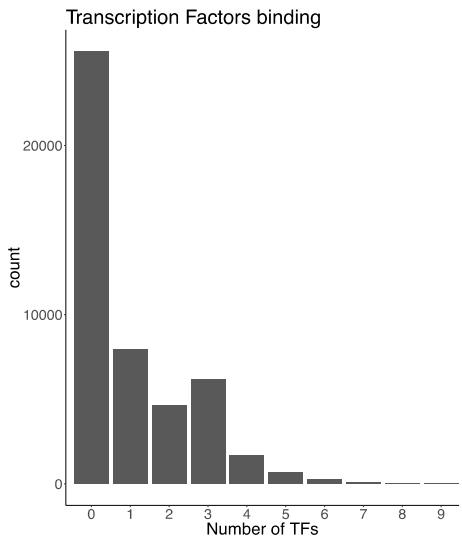


Fig 3.A6: Analysis of TF binding to define binning. Number of overlapping enhancer regions overlapping with 0 or more TFs peaks.

4. RNAPII activation states at enhancer regions mirror their activation

4.1 Introduction

In the previous chapter, I showed that RNAPII is present at enhancer regions, especially concurrently with TFs. RNAPII was previously described to be present at enhancer regions (De Santa et al., 2010) and to transcribe enhancer RNAs (eRNAs) (Andersson et al., 2014a; Arner et al., 2015; Kim et al., 2010; Kaikkonen et al., 2013). eRNAs are currently investigated in the field for their possible role in the regulation of target gene expression. For example, it has been shown that knockout of eRNA with interference techniques can decrease the expression of the target genes (Li et al., 2013;Kaikkonen et al., 2013; Wang et al., 2011). eRNAs may also be involved in chromatin looping (Plank and Dean, 2014), and in promoter proximal pausing release (Schaukowitch et al., 2014a). However, eRNA depletion techniques used in these studies have often interfered with transcription at regulatory regions (Kaikkonen et al., 2013; Plank and Dean, 2014), making it difficult to decouple the effects of transcription from direct effects of the eRNAs. In this regard, it was found that the premature arrest of transcription via insertion of a polyA-cassette at enhancers influenced eRNA production but not transcription itself. This experiment elegantly showed that inhibition of eRNA expression does not have a direct effect on target gene expression (Engreitz et al., 2016). Therefore, more attention has recently been given to the mechanism of transcription at enhancers. Some studies have dealt with RNAPII characterization at regulatory regions, showing for example that RNAPII can bind enhancer regions after their activation through stimuli in macrophages (De Santa et al., 2010) or enhancer activation (Cruz-Molina et al., 2017). It was also shown that extremely active enhancers – super enhancers – are enriched in RNAPII binding (Hnisz et al., 2013). However, the field lacks a detailed investigation of the states of activation of RNAPII at regulatory regions and how these states relate with enhancer states.

Enhancers have also been shown to exist in different state of activation which are linked with their chromatin environment. For example, H3K27ac enrichment correlates with enhancer activity (Creyghton et al., 2010) at active enhancers. Levels of H3K27ac enrichment are used to

distinguish between normal enhancers and super enhancers, which are greatly enriched for H3K27ac and Brd4(Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). Other marks highlight further differences between classes of enhancers, such as P300, that can distinguish between primed enhancers and enhancers transitioning to activation (Zentner et al., 2011), or H3K4me1, which is not indicative of enhancer state and are ubiquitously present at regulatory regions, from poised, to primed to active. Therefore, different chromatin occupancy features may be indicative of enhancer state, and potentially help distinguish the degree of activation of enhancers and their target genes.

4.1.1 RNAPII modifications and gene states

RNAPII states of activation are deeply characterized at coding genes. RNAPII can be post-translationally modified on its C-terminal domain (CTD) and these modifications act as recruitment platform for chromatin remodellers and RNA maturation machineries. These in turn shape the chromatin environment and the state of maturation of RNA at each transcription event. RNAPII is recruited at promoters by TFs. It binds in an unphosphorylated form and it is sub-sequentially modified in its Serine residues. First, the phosphorylation of Serine-5 residues (S5p) causes the transition to transcription initiation. At this stage, RNAPII pauses and it is released from the promoter only after the phosphorylation in Serine-7 and Serine-2 to starts transcription elongation (Brookes and Pombo, 2009). Therefore, RNAPII can be found in different activation states at genes, which are linked to the state of the gene.

Genes can be in different states, such as inactive, when RNAPII is not bound and no mRNA is expressed, or active, when RNAPII is bound and elongating and fully mature RNA is transcribed. However other gene states exist, such as the Polycomb-repressed state, where RNAPII phosphorylated in Ser5 is present at the gene promoter together with Polycomb repressive complexes. Most Polycomb repressed genes in ESCs exist in a *poised* state, are lowly to not express and are linked with development and signalling (Brookes et al., 2012; Stock et al., 2007) and often encode for TFs important in other lineages in differentiated cells (Ferrai et al., 2017). RNAPII at Polycomb-repressed genes is not recognized by an antibody that binds unpshophorylated S2 (RNAPIIS2u, 8WG16ab), which recognizes RNAPII at promoter of active gene simultaneously recognized by RNAPII-S5p antibody. This suggests that poised RNAPII at Polycomb-repressed genes is in a special configuration and/or with a new, yet to identify modifications that could interfere with the binding of 8WG16 antibody to unmodified S2 residues (Brookes and Pombo, 2012).

Although eRNA transcription has been used to identify regulatory regions (Andersson et al., 2014a; Arner et al., 2015; Henriques et al., 2018; Mikhaylichenko et al., 2018), the state of activation of RNAPII has not yet been investigated at enhancers. It remains to be understood how many regulatory regions are bound by RNAPII, whether RNAPII exists in a different states of activation at enhancers and if enhancers bound by RNAPII have different functions compared to regions not bound by RNAPII.

4.2 Aim of the chapter

The presence of RNAPII at regulatory regions and the reported correlation between transcription at enhancer and at genes raises some interesting questions. Is RNAPII present in different states of activation at enhancers? Do RNAPII activations states inform about enhancer states? To answer these questions, I explored the state of RNAPII at published enhancers and the relation with enhancer activation state. Transcription and its regulation at extragenic regions will be investigated in Chapter 5.

In the current chapter, I consider the previously characterized RNAPII activation states at coding regions, and investigate their presence at enhancer regions previously defined in Whyte *et al.* 2013 (Whyte et al., 2013) (Scheme of the chapter Fig 4.1). Furthermore, I sub classified the Whyte enhancers according to RNAPII state and analyse their chromatin marks. Together, the results in the current chapter show that RNAPII exists in different activation states at regulatory regions, which reflect enhancer activation states and suggest that RNAPII occupancy at extragenic regions can be used as a feature to identify enhancers and infer their state.

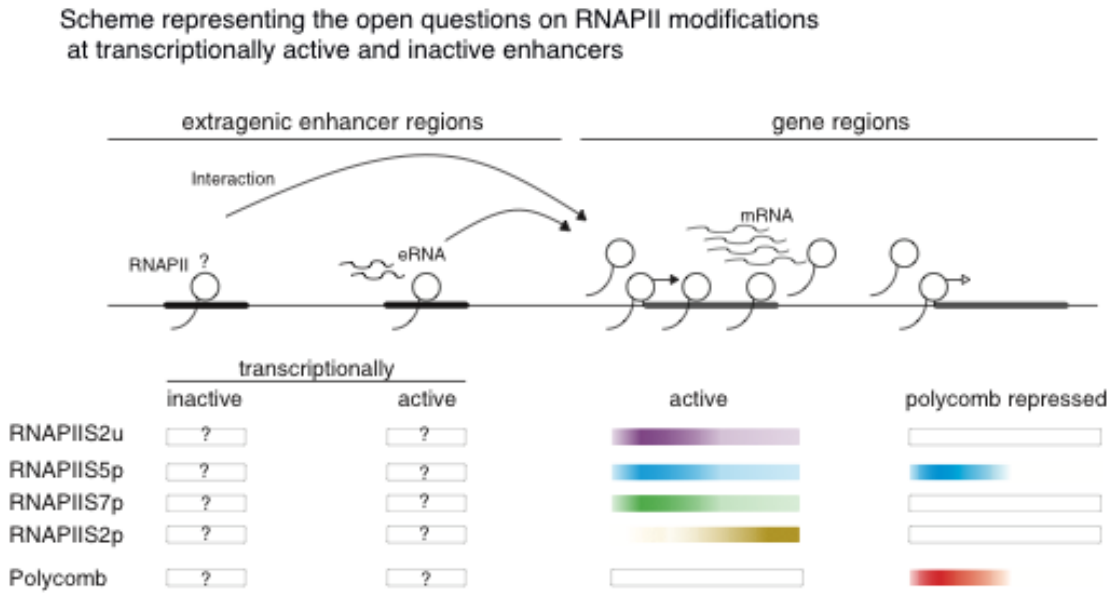


Fig 4.1: Scheme representing the open questions on RNAPII state at transcriptionally active and inactive enhancers compared to active and Polycomb-repressed genes. On the right are represented genes in two states: active and polycomb repressed. Below the bars represent the enrichment of RNAPII modifications and the Polycomb marks H3K27me3 along the gene body. On the left are represented enhancer regions in an active transcribing state and in a not transcribing state. Curved arrows represent chromatin loops. Question marks depict open questions in the field.

4.3 Contribution disclosure

Elena Torlai Triglia processed all the RNAPII datasets and calculated the peaks of occupancy presented in this chapter. Elena Torlai Triglia and Alexander Kukalev processed other additional datasets used in the current chapter, when specified.

4.4 Results

4.4.1 Choice of enhancer lists

In the previous chapter, I showed how heterogeneous is the landscape of enhancers in mESC. I also showed enhancer regions are not only characterized by the presence of specific enhancer marks and TFs, but also often co-occupied by RNAPII. Different RNAPII modifications were present at regulatory regions, which prompted me to explore the state of activation of RNAPII at enhancers. To explore the activation state and role of RNAPII at enhancers, I focused on the Whyte list, where enhancers were defined as co-bound regions of Nanog, Sox2, and Oct4 (Whyte et al., 2013). TFs are understood as the core mediators of enhancer function and known to directly or indirectly recruit RNAPII to chromatin.

Other enhancer lists which were partially analysed were obtained from: Arner *et al.* 2015 (Arner et al., 2015), Chen *et al.* 2012 (Chen et al., 2012), Creyghton *et al.* 2010 (Creyghton et al., 2010), Zetner *et al.* 2011 (Zentner et al., 2011). The results were in general concordant, with some exceptions acknowledged in the text and in the discussion. Cruz Molina and Pradeepa lists are not presented in this chapter, as they were published after the work presented here had been conducted, and not repeated due to time constraints.

4.4.2 RNAPII datasets used in the current chapter

To investigate the presence and state of activation of RNAPII at enhancer regions, I investigated RNAPII best-known post-translational modifications, namely phosphorylation of Serine residues 2, 5 and 7 present in RNAPII's CTD (abbreviated here as S2p, S5p and S7p). Post-translational phosphorylation of RNAPII's CTD has been previously studied in detail and characterize RNAPII's state of activation during initiation, pausing and productive elongation at coding regions (Brookes et al., 2012; Ferrai et al., 2017). I analysed S5p, mark of transcription initiation, S7p, mark of transition from initiation to productive elongation, and Ser2p, mark of fully elongating RNAPII. S5p is also found at poised genes together with the Polycomb marks H3K27me3 and H2AK119ub1 (Brookes et al., 2012; Ferrai et al., 2017; Stock et al., 2007). I also analysed the presence of RNAPII complexes recognised by the 8WG16 antibody, which has a preference for unphosphorylated Serine 2 residues (called here RNAPIIS2u) (Fig 4.2a). The 8WG16 antibody identifies RNAPII present at the promoters of active genes, and it does not bind to the poised RNAPII complexes found at Polycomb repressed genes (Brookes et al., 2012; Ferrai et al., 2017; Stock et al., 2007). These four RNAPII datasets for different post-

translational modification states enable the identification of different states of RNAPII activation at regulatory regions: from poised, to initiating, to fully elongating RNAPII.

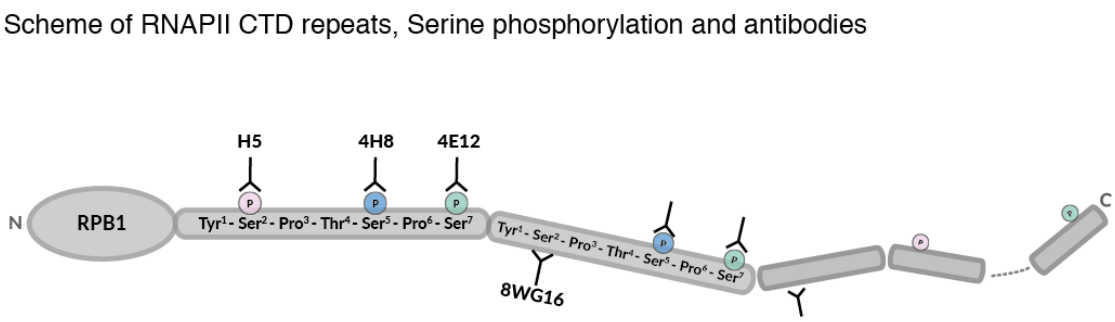


Fig 4.2: RNAPII datasets used in the current chapter. Scheme representing RNAPII-CTD phosphorylation and the antibody used in the current chapter. RBP1 subunit is depicted together with the CTD, composed of heptapeptides repeats. Circles with P represent phosphorilations. Ab clones are specified in the drawing.

Table 4.1 summarizes the RNAPII ChIP-seq datasets used here. Data was downloaded from GEO, remapped for consistency and duplicated reads were removed. To identify the genomic regions positively occupied by each RNAPII modification, datasets were analysed with Bayesian Change-Point (BCP) peak finder (Xing et al., 2012), using Mock IP dataset as control, in histone mark (HM) mode. Processed datasets and lists of positive windows (peaks) were available at the onset of this project, and previously processed by Elena Torlai Triglia. BCP is one of the preferred peak finders currently available as it performs well with both narrow and broad chromatin occupancy (Harmanci et al., 2014; Thomas et al., 2017).

Table 4 1: RNAPII datasets used in the current chapter. List of the RNAPII modifications used in the current chapter. RNAPII peaks were calculated by Dr Elena Torlai Triglia.

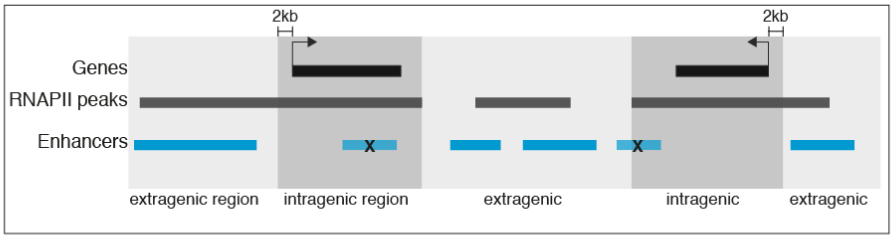
RNAPII datasets	# Peaks	mESC clone	Publication	Antibody	GEO
RNAPII2u	25343	OS25	Brookes et al., 2012	8WG16	GSM850469
RNAPIIS5p	20243	OS25	Brookes et al., 2012	4H8	GSM850467
RNAPIIS7p	14686	OS25	Brookes et al., 2012	4E12	GSM850468
RNAPII2p	7767	OS25	Brookes et al., 2012	H5	GSM850470
Control Mock ChIP	-	OS25	Brookes et al., 2012	-	GSM850473

4.4.3 Strategy to classify extra-genic enhancers according to RNAPII occupancy

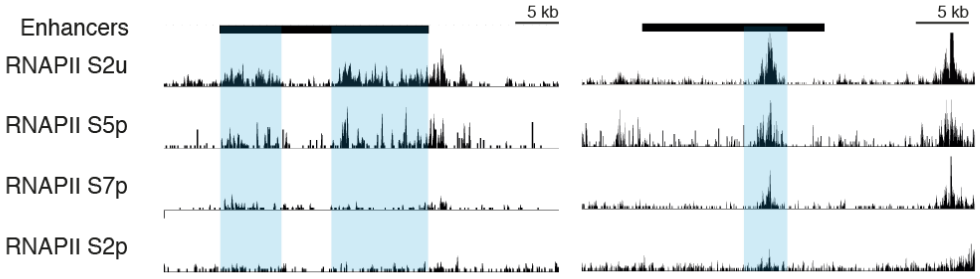
To eliminate the confounding effects of RNAPII roles in genic transcription, I subclassified the Whyte enhancer list in intragenic and extragenic regions. Intragenic regions were defined as the regions inside annotated genes (RefSeq genes (O'Leary et al., 2016)), the 2kb region upstream

from annotated transcription start sites (TSSs) and any RNAPII peaks that contiguously extended beyond annotated coding regions, ie beyond the polyA site, or transcript end site (TES) (Schematic in Fig 4.3a). Regions beyond the TES are excluded to consider the imprecise termination of genic transcription which extends beyond the polyadenylation site (Proudfoot, 2016). In mESCs, I found that RNAPII can readthrough beyond the TES for up to 11kb (For an analysis on how long the RNAPII peaks can extrude the TES please refer to Appendix Fig 4.A1).

a Scheme of extragenic definition



b Examples of long regions with RNAPII peaks



c Scheme of classification approach

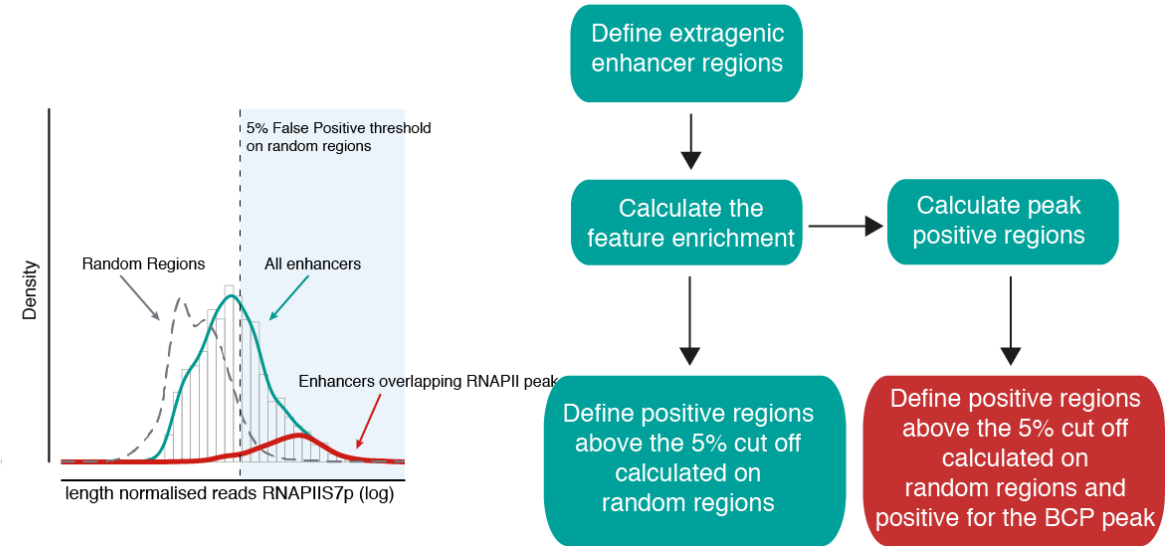


Fig 4.3: Strategy to classify RNAPII modifications at enhancers. a) Scheme representing the strategy to define extragenic enhancers. Enhancer overlapping a gene, in a 2kb window upstream a TSS or overlapping a RNAPII peak transcribing a gene are considered intragenic and are filtered out in this study. Black box: gene, the arrow indicates the TSS and the direction of transcription; grey box: RNAPII peak; blue box: Whyte enhancers. X on top of a blue box signifies that the enhancers is considered intragenic and was not analysed further. Dark shaded regions: intragenic regions. Light shaded regions: extragenic regions. b) Example of long enhancer regions with RNAPII local peaks. ChIP-seq read counts per RNAPII modification are shown. Transparent blue boxes highlight regions inside the enhancer with an RNAPII peak. Image generated with the IGV software. c) Scheme of the classification pipeline. On the left: schematic of the approach used to calculate the extragenic enhancers positive regions. Green line represent the maximal enrichment of ChIP-seq reads at extragenic regions; red line represents the maximal enrichment of ChIP-seq reads at extragenic regions positive for a called peak for the modification; grey dashed line represent maximal enrichment of ChIP-seq reads at 30 randomly permutated extragenic regions. Black vertical dashed line represents the 5% false positive cut-off used to define positive regions. Transparent blue box represent the regions positive for the modification. On the right: steps to classify every regions for each RNAPII modification considered.

The remaining genomic regions were considered as extragenic enhancers, which were therefore confidently annotated for the presence of the RNAPII modification of interest, without the confounding effect of transcription from annotated genes. Of note, annotated lncRNA regions were not used as exclusion regions in this analysis, because some annotated lincRNAs regions were shown to have enhancer function (Orom et al., 2010; Paralkar et al., 2016).

To classify extragenic Whyte enhancers according to the presence of the RNAPII modifications considered here, I calculated the enrichment for each specific modification at the enhancer regions. A visual investigation of Whyte regions on the UCSC genome browser showed that RNAPII peaks overlapping the Whyte enhancers can cover broad regions with non-homogenous RNAPII coverage (Fig 4.3b), which was taken into account in the classification strategy.

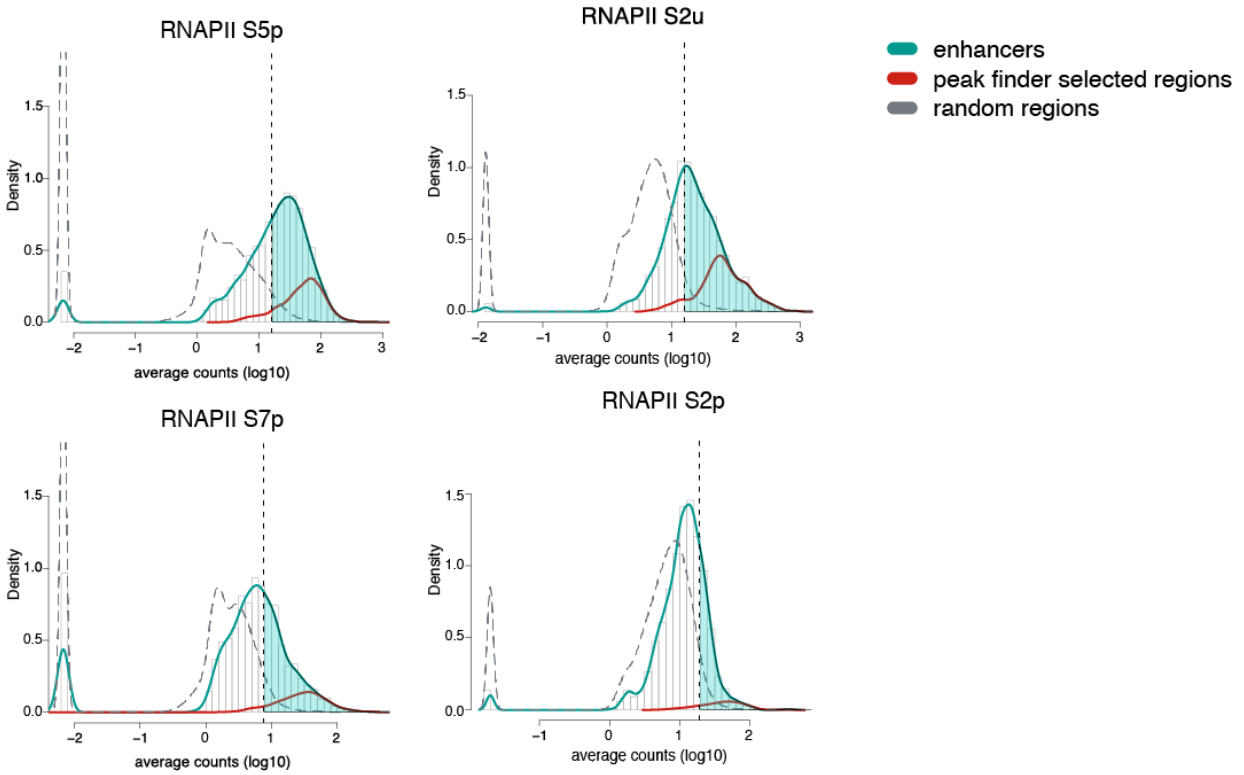
The classification strategy of RNAPII occupancy at enhancers was performed separately for each RNAPII modification as follows (Scheme of the classification approach can be found in Fig 4.3c). First, Whyte enhancers in extragenic regions were identified. Second, extragenic regions were classified as positive or negative for overlap with RNAPII peaks. Third, the enrichment of reads in extragenic enhancer region was calculated. Extragenic enhancer regions longer than 3000bp were divided in bins of 3000bp, and the RNAPII enrichment was calculated per bin and the maximum enrichment was used in the following steps. Fifth, the enrichment for each RNAPII modification was also calculated at randomly shuffled extragenic regions with the same length as the Whyte enhancer regions; shuffling was repeated 30 times. Sixth, the length-normalized region enrichment (total number of reads of the largest enrichment divided by the length of the region) is plotted as a density at both enhancer regions and the random regions. The densities of all random regions are calculated singularly and merged for plotting to improve the statistical power of the random shuffling. Seventh, a 5% False Positive cut-off is calculated considering the top 5% random regions enrichment. All enhancer regions with length-normalized enrichment above the threshold were defined as positive for the specific modification. Steps from 3 to 6 were applied both to all extragenic enhancer regions and to extragenic enhancer regions positive for overlap with RNAPII peaks calculated in the second step. Different cut-off

settings were tested and 5% False Positive cut-off was chosen as it yielded a reasonable number of RNAPII-positive enhancers, while maintaining a low number of positive random regions. After these steps, the extragenic Whyte enhancer regions were classified as positive for each RNAPII modification in two different ways: a "conservative" classification, where extragenic enhancer regions are considered positive for a RNAPII modification if they overlap with a BCP peak for that RNAPII modification with enrichment over the 5% cut-off; and, a "liberal" classification where extragenic enhancer regions are considered positive for a RNAPII modification if they are enriched over the 5% cut-off, irrespectively of an overlap with a BCP peak.

4.4.4 Classification of Whyte enhancers with RNAPII datasets

Using the approach explained in the previous paragraph, I classified the extragenic Whyte enhancers for RNAPIIS5p, RNAPIIS2u, RNAPIIS7p and RNAPIIS2p datasets (Fig 4.4a). Most enhancer regions (60%) were found to be positive for RNAPIIS5p or RNAPIIS2u, and fewer were positive for Ser7p (37%) and Ser2p (23%). 20% of the time extragenic Whyte enhancers overlapped with an RNAPII BCP peak, however the enrichment was below the 5% threshold calculated (Table with all the numbers per region and modification in Fig 4.4b).

a Density graphs used for classification



b Table with numbers of enhancers classified as positive for each mark

	RNAPII-8WG16	RNAPII-S5p	RNAPII-S7p	RNAPII-S2p
Enhancers above the threshold	1615 (61%)	1572 (60%)	977 (37%)	605 (23%)*
Enhancers below the threshold	1013 (39%)	1056 (40%)	1651 (63%)	2023 (77%)*
Peaks above threshold	733 (91%)	557 (90%)	309 (91%)	97 (83%)§
Peaks below threshold (FN)	69 (9%)	62 (10%)	31 (9%)	20 (17%)§

* percentage of all extragenic enhancers
§ percentage of peak-caller positive enhancers

Fig 4.4: Classification of Whyte enhancer. a) Profiles used for the classification of Whyte enhancers for the four RNAPII modifications considered. The blue line represents the maximum ChIP-seq enrichment at all extragenic Whyte enhancer regions. The red line represents the maximum ChIP-seq enrichment at Whyte extragenic enhancer regions positive for the consider mark. The grey dotted line represent the maximum ChIP-seq enrichment at 30 randomly permuted extragenic regions. The vertical dotted line represent the 5% false positive cut-off used to define positive regions. Transparent blue box represent the regions positive for the modification. b) Table with numbers of regions above and below the computed cut-off, for all enhancers and peak finder (BCP) positive enhancers. In green: highlighted the regions above cut-off among all extragenic Whyte enhancer regions. In red: highlighted the regions above the cut-off among extragenic Whyte enhancer positive for the RNAPII peak.

Before exploring the different state of activation of RNAPII at extragenic enhancers, I wanted to understand in what way the two classifications differed and with which one to proceed further in the analysis.

4.4.5 Differences between liberal and conservative classification approach

To understand the differences between the liberal and the conservative classification approaches, I analysed the regions classified differently on a genome browser (Fig 4.5a). For example (Fig. 4.5a left), the liberal approach detected an enrichment on S2p in an enhancer region, which was missed by the peak caller. This kind of discrepancy is probably due to the fact that the BCP peak caller used on the RNAPII datasets takes the local background into consideration, calculated in this case over the control MockIP dataset. Liberal classification, on the other hand, takes into consideration only the background of random regions in the genome. In another example (Fig 4.5a, right), RNAPII enrichment region can lie outside the enhancer region, while the tail protrudes inside it. In this case, the conservative approach would see the region positive for the mark, while the liberal approach would miss it.

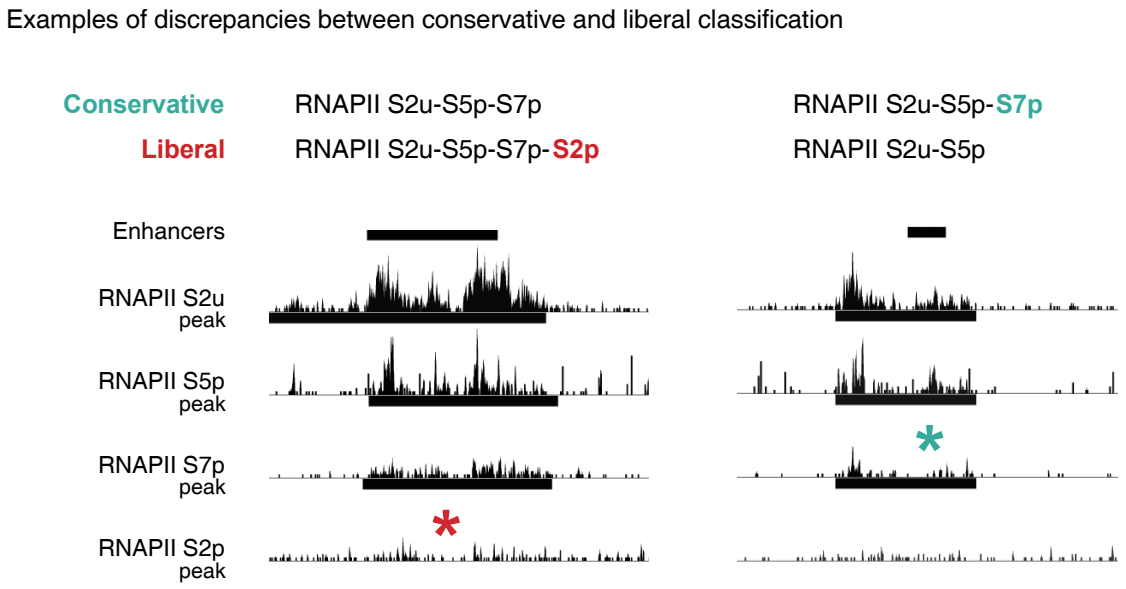


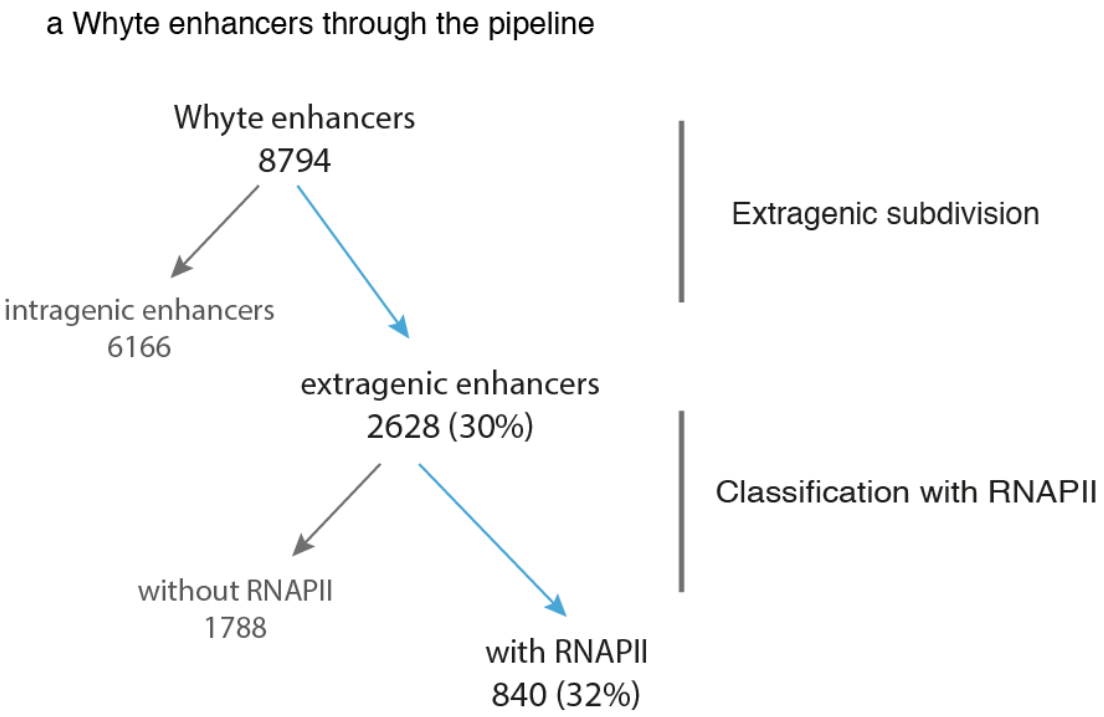
Fig 4.5: Example of discrepancy between liberal and conservative classification. The snapshots show single enhancer regions on the genome browser. On top the classification resulting from the two approaches is specified. In colour: modification present in one classification and not in the other. The star indicates the feature classified differently between the approaches: red star – positive region for that modification in liberal classification and negative in conservative classification; blue star - positive region for that feature in the conservative and negative in the liberal classification. Black bars underneath reads are peaks of the specific feature. Images obtain with UCSC genome browser.

After a thorough investigation of numerous regions with discrepancies between the two classifications, I decided to proceed with the conservative lists. This choice influences the number of positive regions and the statistical power, but gives us confidence on the classification of RNAPII states and the following results. All the analyses presented in the following chapter were also performed on the liberal list and were in general in accordance with the conservative

list. In conclusion, I developed a strategy to categorise enhancer lists into extragenic or intragenic and to classify them for RNAPII modifications.

4.4.6 Whyte enhancers through the RNAPII classification pipeline

The proportion of extragenic enhancers in the Whyte list is 30% of the total (Fig 4.6a), which is consistent with analysis on other lists, such as Chen *et al.* 2012 (Chen et al., 2012)and Arner *et al.* 2015 (Arner et al., 2015)(data not shown). Of the 2628 extragenic Whyte enhancers, 840 (32%) are co-occupied by at least one form of RNAPII, while 1788 (68%) are not bound by any of the RNAPII marks considered here.



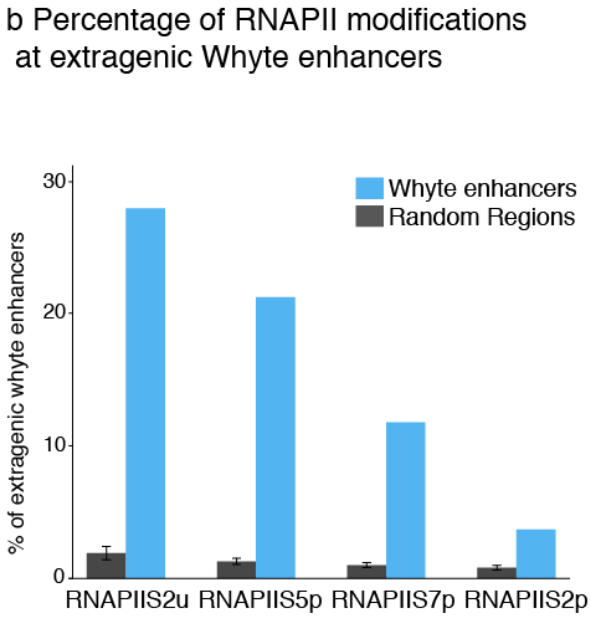


Fig 4.6: Whyte enhancers through the classification pipeline. a) Graph showing the number of Whyte enhancers per step of the pipeline. Whyte enhancers are first subdivided in intragenic and extragenic regions and extragenic regions are classified as positive of negative for RNAII presence. Blue arrows indicate the path to the regions selected to be further investigated in this study. b) Percentage of each RNAPII modification at Whyte extragenic enhancers. On the x-axis: RNAPII modification. On the y-axis: percentage of extragenic Whyte enhancers positive for that modification. Blue bars represent Whyte enhancers. Grey bars represent extragenic permutated regions. Error bars are computed on randomly permuted regions classification and indicate standard deviation.

All RNAPI modifications were from present at extragenic Whyte enhancers above random. The most represented RNAPII dataset at extragenic Whyte enhancers is RNAPIIS2u (28% of extragenic enhancers), followed by (20% of extragenic enhancers), RNAPIIS7p (11% of

extragenic enhancers) and RNAPIIS2p (7% of extragenic enhancers) (Fig 4.6b). Similar proportions were found at extragenic Whyte enhancers classified with the liberal approach (Appendix Fig 4.A2a). After classifying all extragenic Whyte enhancer regions for the presence or absence of RNAPII modifications, I moved forward to classify the different states of RNAPII at extragenic Whyte enhancers.

4.4.7 RNAPII is found in different activation states at extragenic Whyte enhancers

To understand the states of RNAPII at regulatory regions, I proceeded analysing the different combinations of RNAPII modifications found at extragenic Whyte enhancers. I found different RNAPII states, among which the most represented states of RNAPII are: Only RNAPIIS2u (258), RNAPIIS2u-S5p-S7p (198), RNAPIIS2u-S5p (168), Only RNAPIIS5p (96), RNAPIIS2u-S5p-S7p-S2p (83) (Fig 4.7a). Other less abundant categories were not considered further, as they comprise 5% of the total RNAPII-bound extragenic Whyte enhancers. I also briefly explored regions positive for RNAPIIS2u-S7p (20) in the genome browser to understand if they could be added to one of the above mentioned classes, in case they fell just below the 5% FP threshold of RNAPIIS5p detection. However, these regions were very mixed and were therefore not considered further.

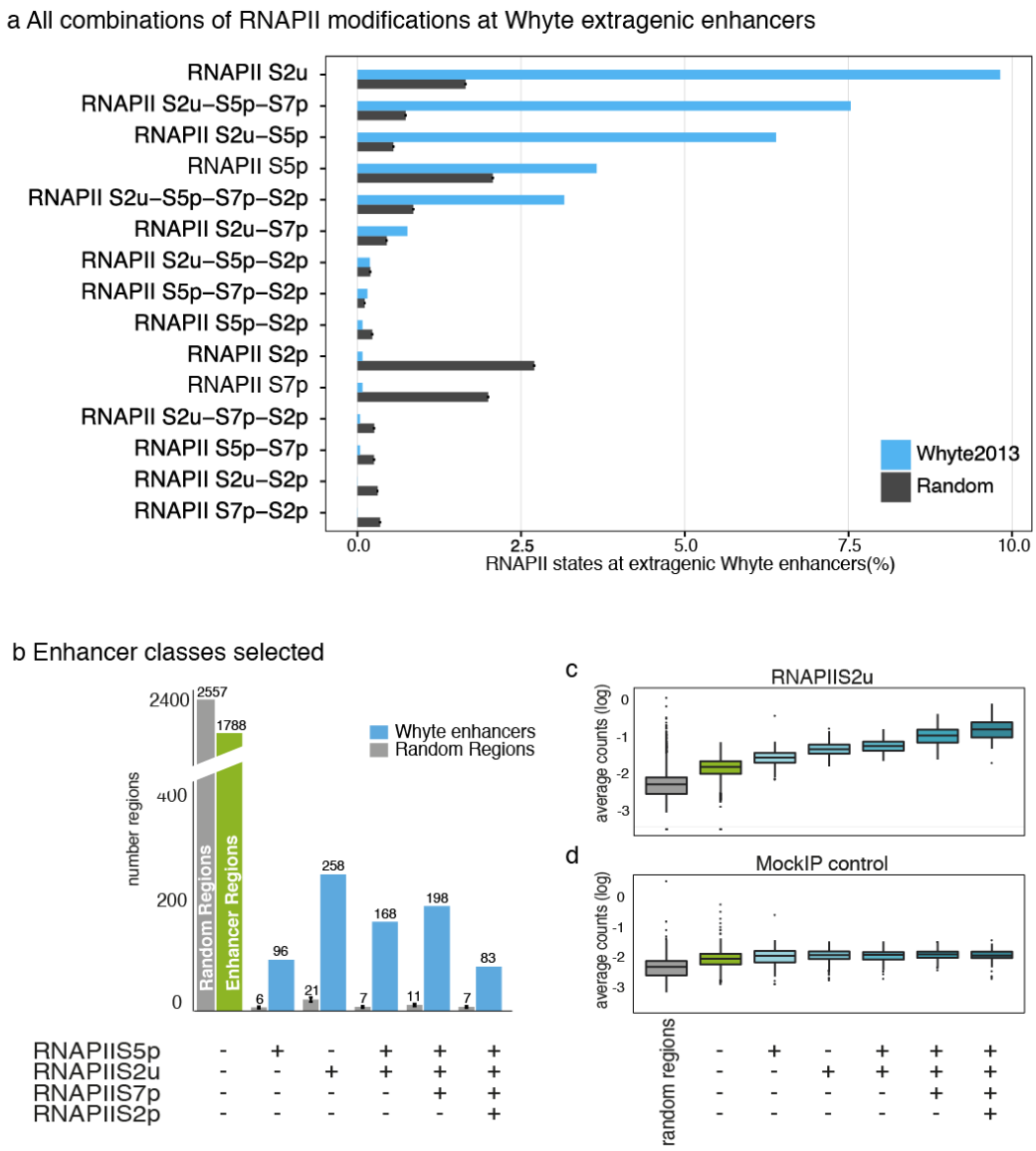


Fig 4.7: RNAPII is present in different states at extragenic Whyte enhancers. a) Plot showing all the combinations of RNAPII modifications at extragenic Whyte enhancers. On the y-axis: RNAPII combinations. On the x-axis: percentage of extragenic Whyte enhancers. Blue bars represent Whyte enhancers. Grey bars represent extragenic permutated regions. Error bars are computed on randomly permuted regions classification and indicate standard deviation. b) Enhancer classes selected for further study. On the x-axis: state of RNAPII. On the y-axis: number of regions. Blue bars represent extragenic Whyte enhancers bound by RNAPII. Green bar represent extragenic Whyte enhancers not bound by RNAPII. Grey bars represent extragenic permutated regions. Error bars are computed on randomly permuted regions classification and indicate standard deviation. c) Enrichment of unphosphorylated Ser2 at RNAPII-bound enhancers, divided by RNAPII state. Average count: number of ChIP-seq reads per region divided by the region length. Log values are shown for clarity. A pseudo count equivalent to the minimum value not equal to 0, divided by 10 was added, to avoid log of 0. Random regions are randomly permuted extragenic regions. d) Enrichment of MockIP at RNAPII-bound enhancers, divided by RNAPII state. Average count: number of ChIP-seq reads per region divided by the region length. Log values are shown for clarity. A pseudo count equivalent to the minimum value not equal to 0, divided by 10 was added, to avoid log of 0. Random regions are randomly permuted extragenic regions.

The RNAPII states selected at extragenic enhancers are similar to the ones described at genes: from RNAPIIS5p, associated with poised genes, to the fully elongating state with RNAPIIS2u-S5p-S7p-S2p. Noticeably, I find also regions marked by RNAPIIS2u alone (Fig 4.7b), which is not a common RNAPII state at genes where only 38/18860 gene promoters were marked by RNAPIIS2u along (Brookes et al., 2012). This result is consistent across different published enhancer lists (Appendix Fig 4.A2b). As a side note, while the classes recovered at other lists are

the same, the proportion between them varies. This is most probably a reflection of the states of the enhancers in each list: Arner enhancers, which are mainly co-occupied by elongation RNAPII or RNAPII transitioning to elongation (Appendix Fig 4.A2b), are defined as transcribed regions outside promoters identified by CAGE tags. These differences between proportions of classes linked with the specificity of the enhancer lists could suggest a relationship between the enhancer state and the RNAPII state.

To understand whether the state of activation of RNAPII is also linked with increased RNAPII recruitment at extragenic enhancer regions, I analysed the enrichment of RNAPIIS2u. RNAPII-bound enhancers are progressively more enriched for RNAPIIS2u, the more active is the state of RNAPII (Fig 4.7c). Importantly, this gradual increase in RNAPIIS2u occupancy is not reflected in the Mock-IP datasets, suggesting that it is not a secondary effect of open chromatin (Fig 4.7d), but connected with the state of RNAPII at extragenic enhancers.

Taken together, the results show that RNAPII binds in different states of activation to a subset of extragenic enhancers, and these states resemble the ones previously descried at genes, suggesting that RNAPII transcription at enhancers may be controlled by similar regulatory mechanisms as at genes. Moreover, RNAPII binds enhancers at increasing levels concordant with RNAPII activation state.

4.4.8 RNAPII-bound enhancers are associated with early development genes and negative regulators of differentiation

To investigate differences between extragenic Whyte enhancers bound or not bound by RNAPII, I first analysed their putative target genes. As a first approximation, I used the GREAT software (McLean et al., 2010), which performs Gene Ontology (GO) analysis based on the closest genes to the enhancers of reference. This is an approximation, as enhancer targets can be found very far away from their target genes with other genes occurring between them (Arner et al., 2015; Lettice et al., 2018). Interestingly, enhancers bound by RNAPII and enhancers not bound by RNAPII show a different enrichment in biological process' GO terms. Enhancers bound by RNAPII have more terms of related to stem cells maintenance and negative regulators of differentiation, while enhancers without RNAPII are more associated with genes related with positive regulation of differentiation, and therefore with genes that might be primed for activation but not yet active (Fig 4.8a).

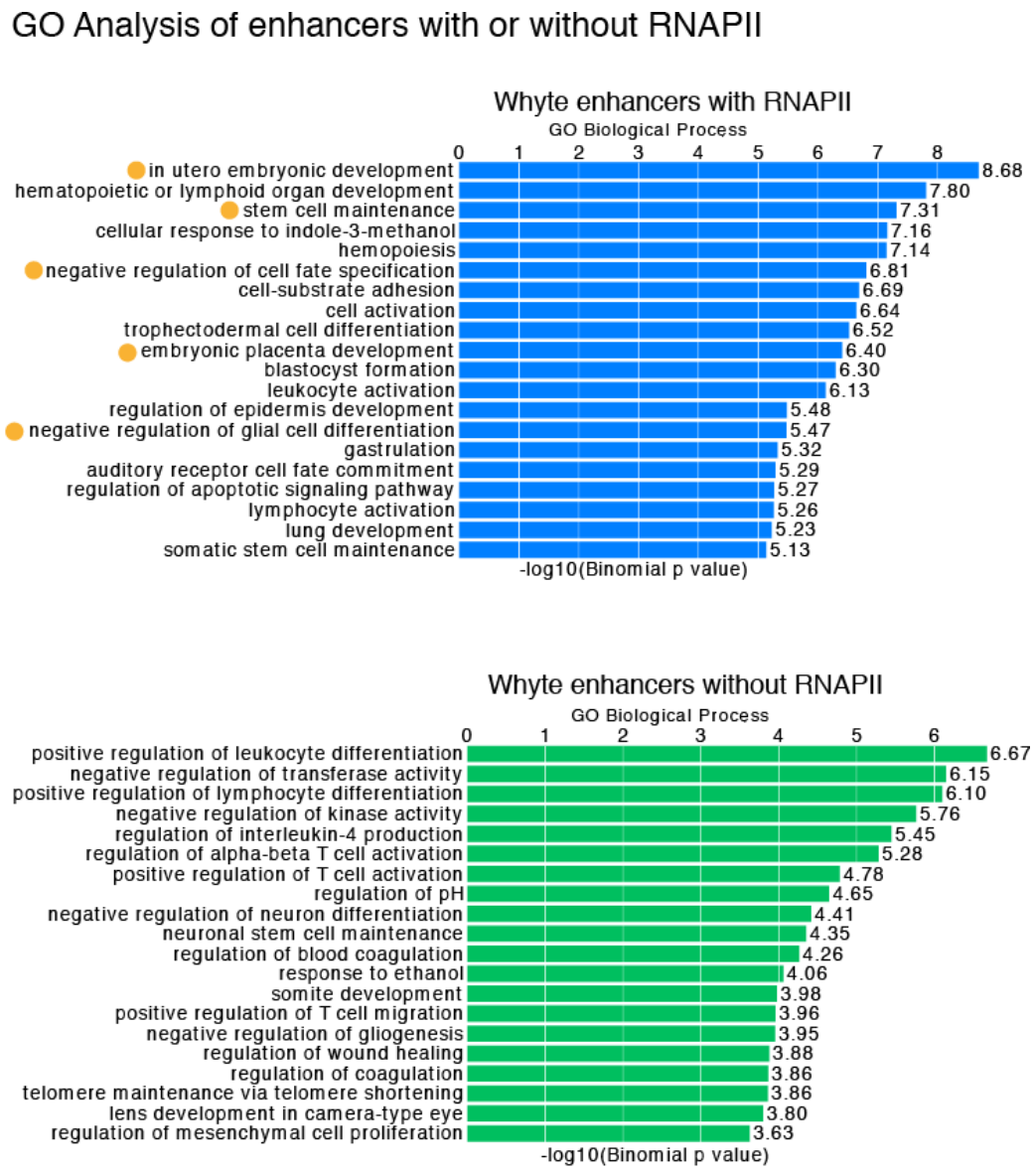


Fig 4.8: RNAPII bound enhancers are associated more with genes involved in stem cell maintenance. GO analysis at RNAPII-bound and RNAPII not-bound enhancers. On top in blue: enhancers bound by RNAPII. On bottom in green: enhancers not bound by RNAPII. Analysis performed with the GREAT software with standard parameters and whole genome as background. Yellow circles: highlighted terms that refer to stem cells and negative regulator of differentiation.

This analysis highlights differences wbetween enhancers bound by RNAPII and enhancers not bound by RNAPII which are likely to have biological relevance, and which may allow more robust categorization of functional enhancers in a give cell type or differentiation stage. It suggests that enhancers bound by RNAPII could be more active in mESC than the ones not bound by RNAPII, but still occupied by Sox2, Nanog, and Oct4.

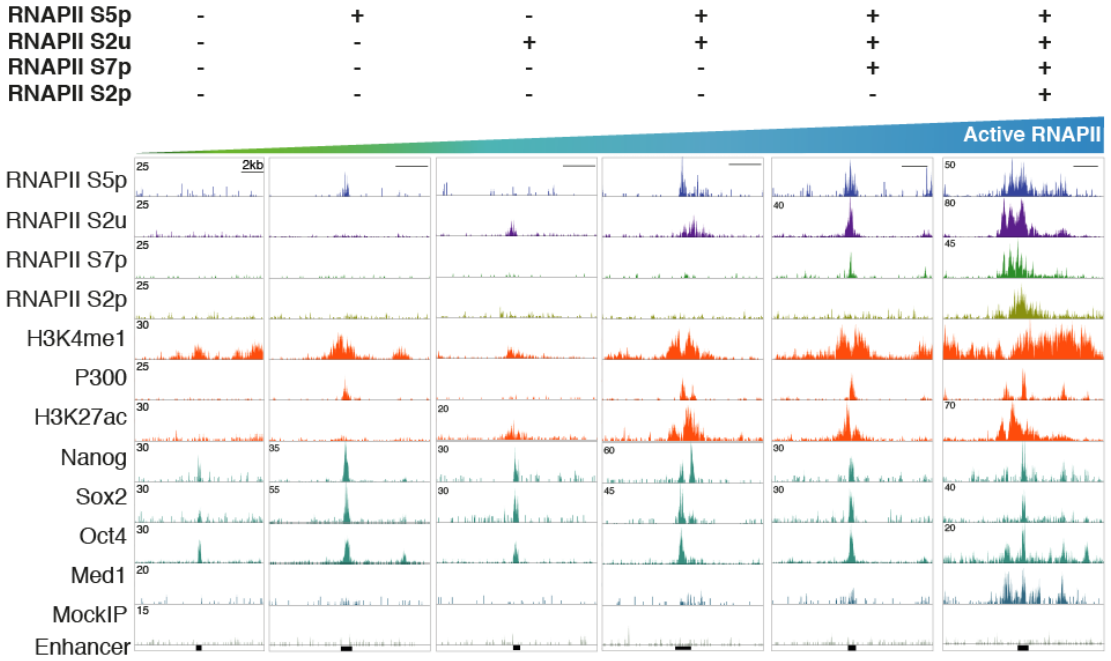
4.4.9 RNAPII-bound enhancers have diverse features related with RNAPII state

To explore whether specific classes of RNAPII-bound enhancers might show different properties, I first inspected single enhancer regions in the genome browser (Fig 4.9a). Extragenic

Whyte enhancers are bound by TFs, as expected, and show a broad enrichment for H3K4me1, which marks all classes of enhancers. Other accepted enhancer marks are present to different extents. For example, the more RNAPII is active (i.e. marked not only by S5p but also S7p and S2p), the more H3K27ac, Med1, and P300 occupancy increases.

Figure 4.9

a Examples of classes RNAPII-bound enhancers



b Positional enrichment of RNAPII modifications per class of RNAPII-bound enhancers

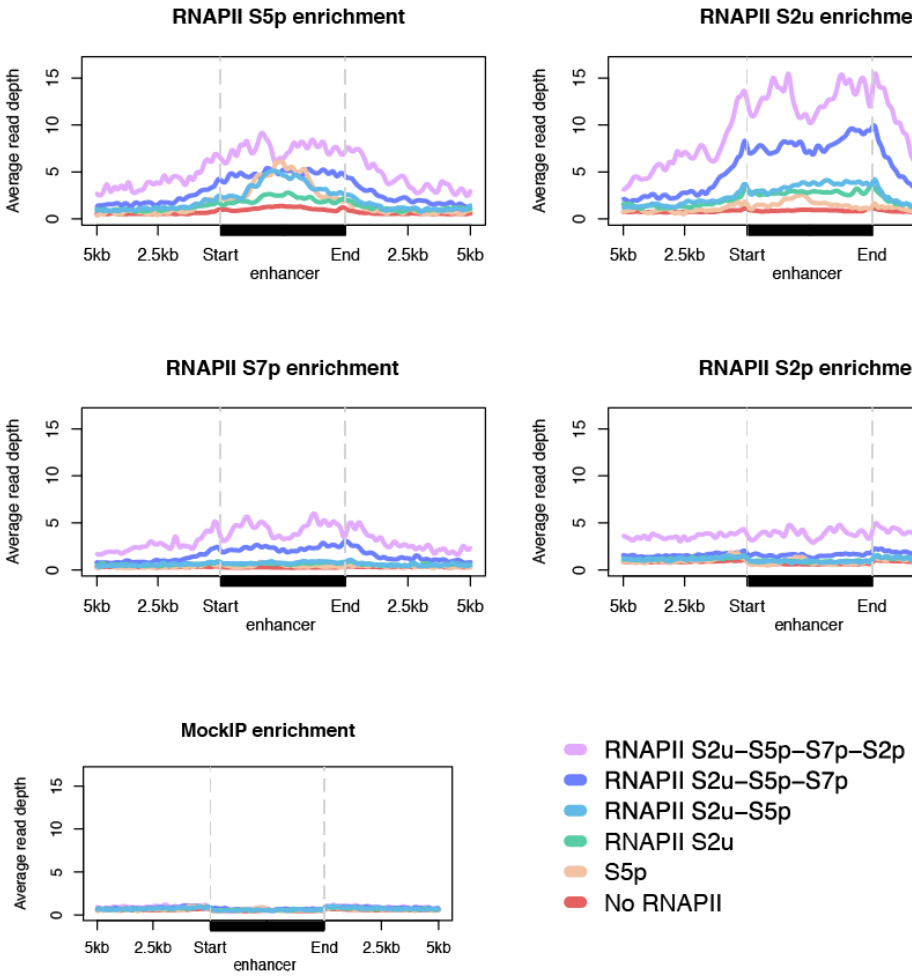


Fig 4.9: RNAPII-bound enhancers show diverse features. a) UCSC tracks of single Whyte extragenic enhancers classified for RNAPII state. Enhancer classes are order from enhancers not bound by RNAPII to enhancers bound by the most active RNAPII. Combinations of the RNAPII states are indicated on the top. On the left: name of the mark of the track. Black boxes on the bottom are enhancer regions. b) Average depth per nucleotide of RNAPII modification enrichment at different classes of RNAPII-bound enhancers and their surroundings. On top of each graph: ChIP-seq dataset analysed. Every line represents the average read count per class of enhancer.

To quantify the enrichment of each RNAPII modification at each group of enhancer classified according to the presence or absence of different RNAPII modifications, I next calculated the average enrichment after normalizing enhancer length. Enrichment of RNAPII per region was calculated at the nucleotide levels and averaged among the classes. Interestingly, RNAPII levels of occupancy vary across classes, and RNAPII is increasingly found to expand beyond the Whyte enhancer coordinates the more active its state is (Fig 4.9b). For example, enhancers marked by all RNAPII marks studied here, also show higher enrichment in S5p and S2p across all classes, and S2p clearly expands beyond the coordinates of the Whyte enhancers, possibly indicating a transcribing RNAPII. This trend is especially evident for RNAPIIS5p and RNAPIIS2u, but also holds true for RNAPIIS7p and RNAPIIS2p. Next, I investigated the length of enhancers marked by RNAPII modifications increasingly associated with productive transcription (Fig 4.10a), and found that RNAPII-bound enhancers tend to be longer than unbound ones, and interestingly that Whyte enhancers associated with fully activated RNAPII (S2p and/or S7p) tend to be the longest. To test whether these differences in RNAPII activation and enhancer window length were also reflected in the genomic location of extragenic enhancer classes relative to candidate target genes, I analysed their distance to the closest gene. I found that enhancers are located both upstream and downstream the closest genes with no noticeable differences per class (not shown), with the majority separated by around 10-50kb from the nearest gene (333/803) (Fig 4.10b). There is a 2-5kb tendency for enhancers associated with elongating forms of RNAPII to be closer to their target genes. Interestingly, enhancers can be located up to 100kb from the closest gene and this feature is common for all the differently classified enhancers (173/803). On the genomic level, different classes of enhancers do not show any obvious position difference (Appendix Fig 4.A3).

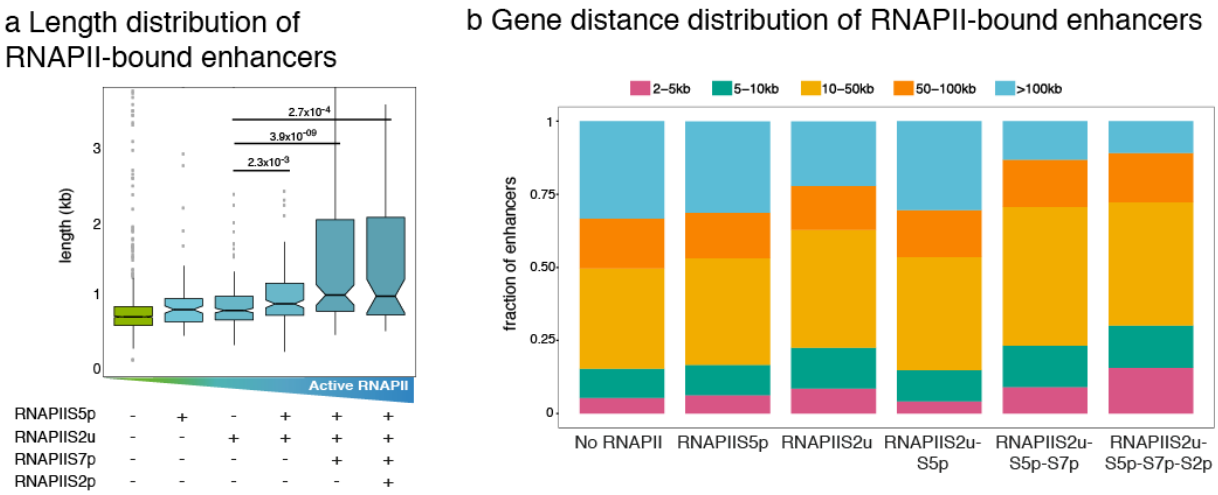


Fig 4.10: Features of RNAPII-bound extragenic Whyte enhancers. a) Length distribution of extragenic Whyte enhancers divided by RNAPII class. On x-axis: combinations of RNAPII modification at enhancers specified. On the y-axis: length, expressed in kilo bases. p-value calculated with Wilcoxon test. b) Gene distance distribution of extragenic Whyte enhancers. On bottom enhancer classes are specified. Distances are binned in: 2-5kb from the closest gene; 5-10kb from the closest gene; 10-50kb from the closest gene; 50-100kb from the closest gene; >100 kb from the closest gene. On the y-axis: fraction of enhancers in bin.

Taken together, these results show that subdividing enhancers according to RNAPII activation state provides further hints of their biological roles, that these classes have specific characteristics, in terms of enhancer length and occupancy. The observation that RNAPII-bound enhancers tend to be longer with increased RNAPII activation state with higher RNAPII enrichment suggest that these longer enhancer regions may build a specific chromatin environment.

4.4.10 RNAPII activation state at enhancers reflects their activation states defined by TF and histone marks

To test whether RNAPII-bound enhancers have diverse enrichment for enhancer features and whether these differences are associated with specific RNAPII activation states, I first analysed the presence of features commonly accepted to mark enhancer regions. I considered: Nanog, Sox2 and Oct4, which are the mESC TFs used in defining enhancers in the Whyte list; H3K4me1, which marks enhancers irrespectively of activation state (primed, poised and active); P300, which is the enzyme responsible for H3K27 acetylation and marks primed enhancers converting to active and; H3K27ac, whose levels are positively associated with enhancer activity (Creyghton et al., 2010; Hnisz et al., 2013). Heatmap representation of TF, H3K4me1, p300 and H3K27ac ChIP-seq occupancy centred on length normalized extragenic Whyte enhancer regions (Fig 4.11a) show that transcription factors are enriched at all extragenic Whyte enhancer regions, independently of RNAPII presence, with a slight preference for RNAPII-bound enhancer regions (Fig 4.11b) TF binding is detected preferentially inside the defined enhancer regions, as expected from their original definition (Whyte et al., 2013). Interestingly, RNAPIIS2u has the lowest

enrichment for Nanog and Oct4 among the RNAPII-bound classes, while comparable levels of Oct4. This difference may be due to Oct4 binding before the other factors to prime the region and could suggest that these enhancers are less active compared to other RNAPII-bound enhancers. In contrast, H3K4me1 spreads over the enhancer region, and is enriched at their boundaries, with lowest overall enrichment at enhancers not bound by RNAPII and increased enrichment with RNAPII activation state. These observations suggest that RNAPII presence may help distinguish enhancer activation states. P300 and H3K27ac, on the other hand, show a preferential enrichment for RNAPII-bound enhancers (Fig 4.11a,c). RNAPII states associated with productive transcription at active genes (RNAPIIS2u-S5p, RNAPIIS2u-S5p-S7p, RNAPIIS2u-S5p-S7p-S2p) are significantly more often enriched for P300 than less active configurations of RNAPII (RNAPIIS2u or RNAPIIS5p alone).

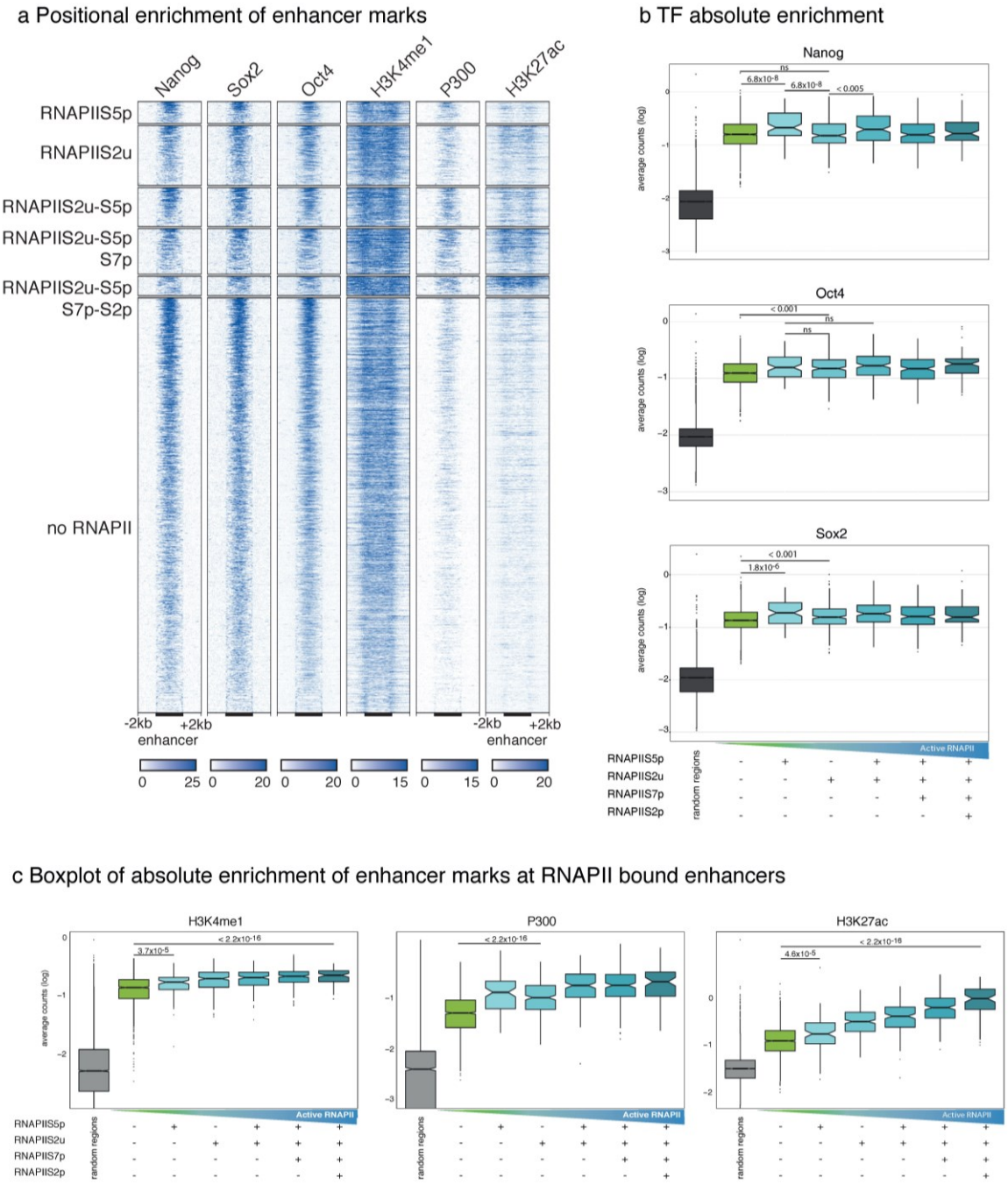


Fig 4 .11: RNAPII state at extragenic enhancers reflect their activation state. a) Positional heatmaps of TFs and enhancer marks at different classes of RNAPII-bound enhancers and +/- 2kb. Darkblue: stronger enrichment, white: lower enrichment. Heatmap generated with deeptools. b) Absolute enrichment of TFs at enhancer regions divided by RNAPII state. Average count: number of ChIP-seq reads per region divided by the region length. On the x-axis: RNAPII modification combiantions. On the y-axis: average count of ChIP-seq reads. Log values are shown for clarity. A pseudo count equivalent to the minimum value not equal to 0, divided by 10 was added, to avoid log of 0. Random regions are randomly permuted extragenic regions. p-value calculated with Wilcoxon test. c) Absolute enrichment of enhancer features at enhancer regions divided by RNAPII state. Average count: number of ChIP-seq reads per region divided by the region length. On the x-axis: RNAPII modification combinations. On the y-axis: average count of ChIP-seq reads. Log values are shown for clarity. A pseudo count equivalent to the minimum value not equal to 0, divided by 10 was added, to avoid log of 0. Random regions are randomly permuted extragenic regions. p-value calculated with Wilcoxon test.

Taken together these results show how RNAPII-bound enhancers have specific properties, which are linked with enhancer states. The incremental occupancy of H3K27ac and the differential enrichment of P300 strongly suggest that the state of activation of RNAPII mirrors the state of

activation of enhancers. TF binding alone would not be enough to clearly distinguish these properties that are captured by RNAPII states. Enhancers bound by RNAPII show typical enhancers marks. Importantly, marks associated with enhancer activation states are differentially enriched at enhancers with more active RNAPII, suggesting that not only the presence of RNAPII, but also its state of activation, is linked with the enhancer activation state.

4.4.11 RNAPII-bound enhancers show diverse feature enrichment for RNAPII state

In the previous section, I found that RNAPII states at enhancers are linked with enhancer state, based on well-established enhancer chromatin marks. To understand whether RNAPII state at extragenic Whyte enhancers is linked with other features, I analysed different factors and histone marks. RNAPII-bound enhancers have higher enrichments for marks linked with active chromatin, such as H3K9ac and TFs, compared to unbound enhancers (Fig 4.12a). However, other features are similarly enriched at Whyte enhancers irrespectively of RNAPII occupancy and activation state; for example CTCF, H3K36me3, and repressive marks such as H3K9me3 and H3K27me3, are not specifically enriched. Smad1, instead, shows an interesting trend, being more enriched at regions marked by RNAPIIS5p alone, but not the other RNAPII marks associated with productive transcript.

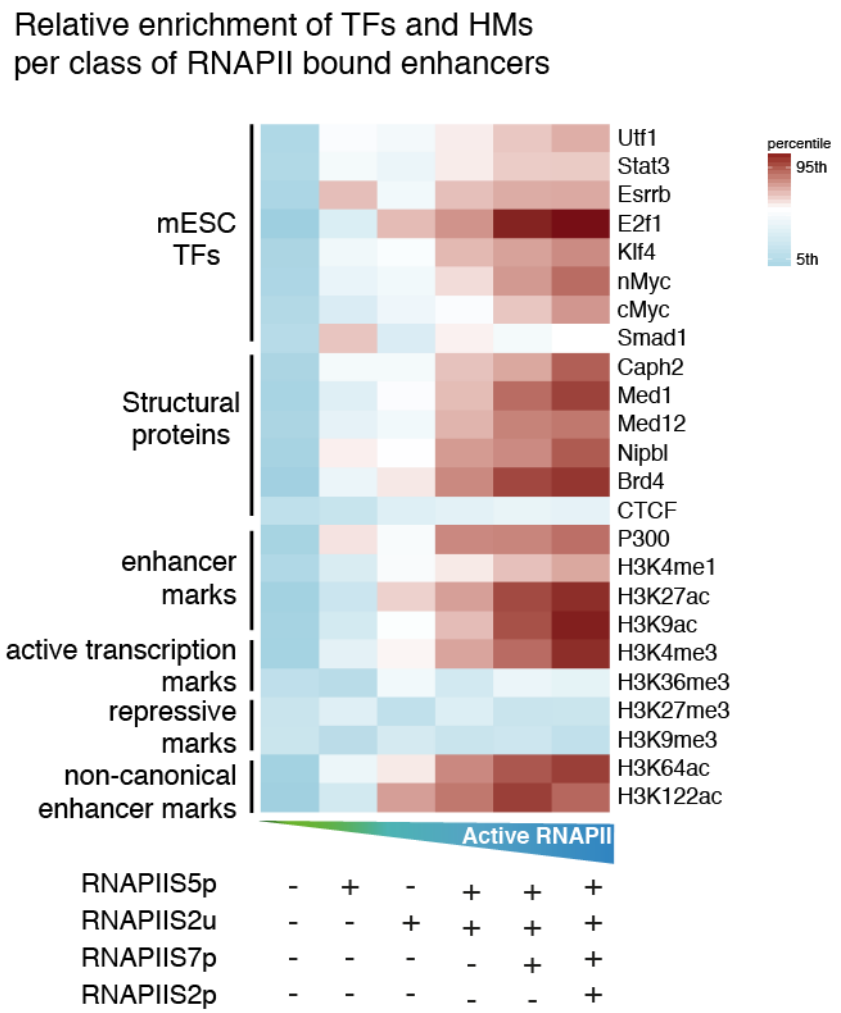


Fig 4.12: RNAPII-bound regions show diverse feature enrichment. Heatmap showing relative average enrichment (percentile of enrichment) per class of RNAPII-bound enhancers of HMs, TFs, and structural proteins at RNAPII bound enhancers. Relative average enrichment was calculated as the mean of the normalised enrichment per class, for comparison purposes. Red: higher relative enrichment; light blue: lower relative enrichment. Classes of enhancers are specified at the bottom of the graph.

Active RNAPII states at extragenic enhancers are associated with increase TF occupancy, such as E2f1 and Klf4. Interestingly, factors connected with chromatin architecture, Med1, Caph2, Nipbl, are also increasingly enriched at extragenic Whyte enhancers with increased RNAPII activation state. This could be indicative of an involvement of these regions in chromatin contacts. Interestingly, CTCF is depleted at these regions.

It is also noteworthy that the enhancers marked only by RNAPIIS2u and RNAPIIS5p classes show different behaviours from the enhancer regions associated with increasing states of RNAPII activation. Not only P300 is differently enriched between RNAPIIS2u and RNAPPIIS5p compared to the other RNAPII-bound classes of extragenic enhancers, but also other factors show this behaviour, such as Brd4 and Caph2. This suggests again the existence of differences between enhancers bound by primed or active RNAPII configurations.

In conclusion, these analyses show that RNAPII is present at different states of activation at enhancers, which mirror the enhancer activation state. Moreover, the state of the RNAPII modification dissects the chromatin state of enhancer activation, a distinction missed by Nanog, Sox2, and Oct4 occupancy.

4.4.12 Active states of RNAPII are found at super enhancer regions

The Whyte enhancer list contains normal enhancers and super-enhancers (SE), which were shown to be the most active among all the enhancers in a cell type (Hnisz et al., 2013; Whyte et al., 2013). SE comprise only a small proportion of the extragenic enhancer regions selected (40/2628). To understand whether SE have a preferential RNAPII state, I investigated them separately. First, I have classified them for RNAPII state, and I have analysed the enrichment for enhancer marks at these regions.

SEs are associated with active states of RNAPII, with 77% (31/40) of the SE regions bound by RNAPII either in fully elongating, RNAPIIS2p form, or transitioning to elongating, and marked by RNAPIIS7p (Fig 4.13a). Only 2 of the extragenic SEs are not bound by RNAPII. Interestingly, RNAPIIS5p alone is not found at SEs, which may imply that SEs are not in a poised or repressed state. Additionally, SE are enriched in RNAPII and enhancer marks (Fig 4.13b), which is in accordance with the previous results in the current chapter, were regions with active RNAPII are more enriched for active chromatin marks.

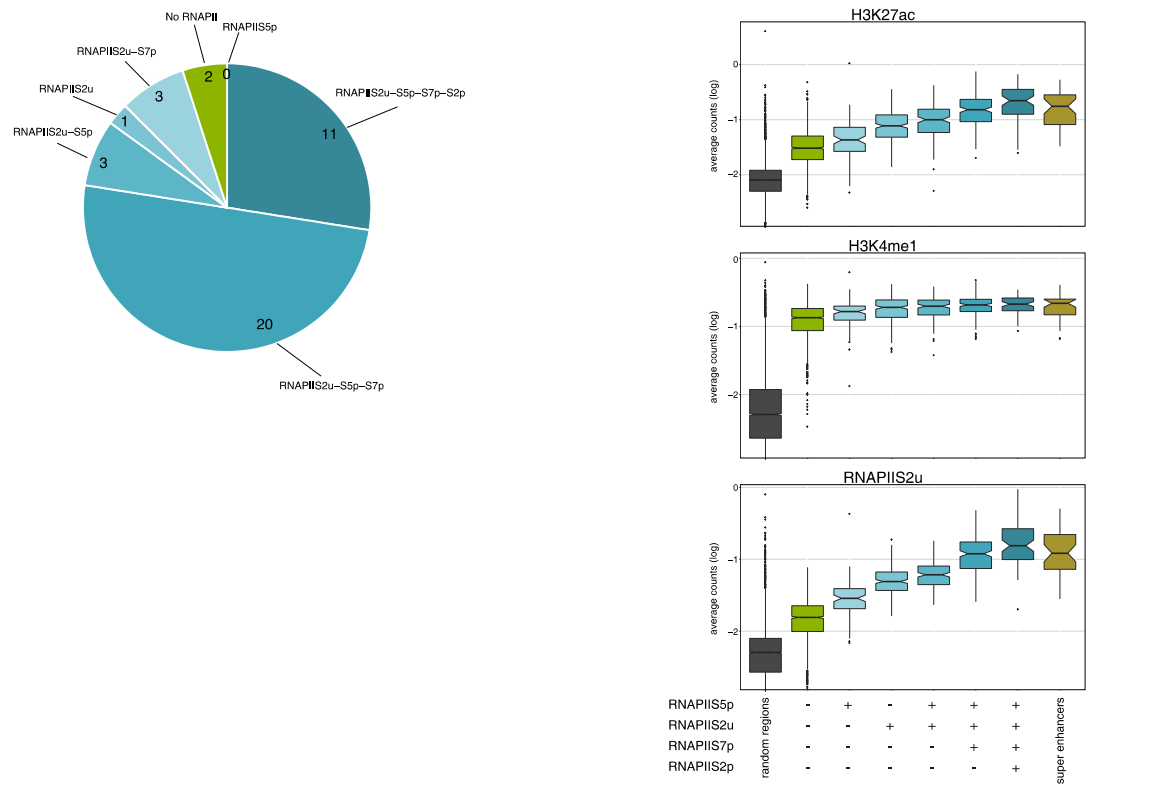


Fig 4.13: Super enhancers are bound by active RNAPII. a) Pie chart showing the proportion of extragenic SE classified with different forms of RNAPII. Numbers of extragenic SE per class are indicated. b) Absolute enrichment RNAPIIS2u, H3K4me1, H3K27ac at enhancer regions divided by RNAPII state and SE. Average count: number of ChIP-seq reads per region divided by the region length. On the x-axis: RNAPII modification combinations. On the y-axis: average count of ChIP-seq reads. Log values are shown for clarity. A pseudo count equivalent to the minimum value not equal to 0, divided by 10 was added, to avoid log of 0. Random regions are randomly permuted extragenic regions. p-value calculated with Wilcoxon test.

4.4.13 RNAPII occupancy distinguishes enhancers and super enhancers

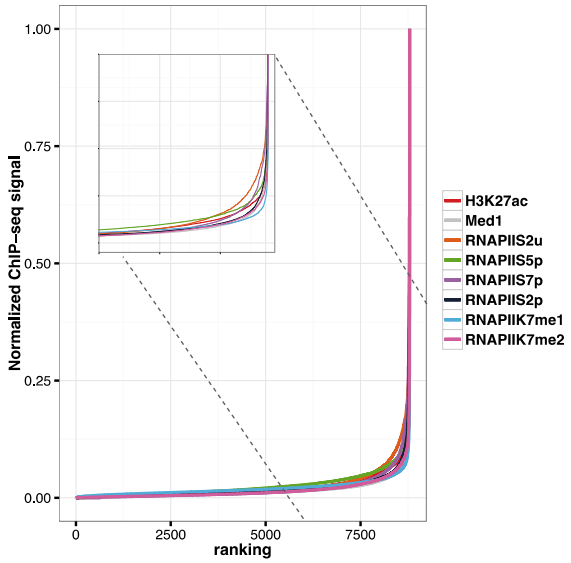
Whyte *et al.* 2013 (Whyte et al., 2013) and Hnisz *et al.* 2013 (Hnisz et al., 2013) showed that Med1 and H3K27ac are more enriched at SEs compared to normal enhancers and their enrichment can be used as a proxy to identify SEs. To test whether RNAPII occupancy alone can distinguish SEs from normal enhancers, I compared ChIP-seq enrichment of all RNAPII forms considered, as well as RNAPIIK7me1 and RNAPIIK7me2, in all the enhancer regions ranked according to the same enrichment. RNAPIIK7me1 and RNAPIIK7me2 were recently described in the lab to mark transcription initiation (Dias et al., 2015).

Interestingly, I found that RNAPII modifications show a similar trend to that of Med1 and H3K27ac (Fig 4.14a). Direct comparison with the position of the original Whyte SE list show that RNAPII occupancy performs as well as Med1 and H3K27ac (Fig 4.14b, red dots indicate originally identified SE). This analysis was conducted over the total set of Whyte enhancers,

including both extragenic and intragenic. It is remarkable that even in this case, RNAPII datasets show good performance; this was especially true for RNAPII-K7me1 and me2 (not shown and Fig. 4.14b, respectively). For comparison, Fig. 4.14b also shows Med1, used in Whyte *et al.* 2013 (Whyte et al., 2013) to define super enhancers and H3K27ac, which shows similar ranking but does not exactly recapitulate Med1.

This finding confirms that RNAPII states are linked with enhancer states and suggests that RNAPII modifications can be used to differentiate between normal and SEs.

a Ranking strategy to define SE with various datasets



b Performarnce of selected features in enhancer ranking

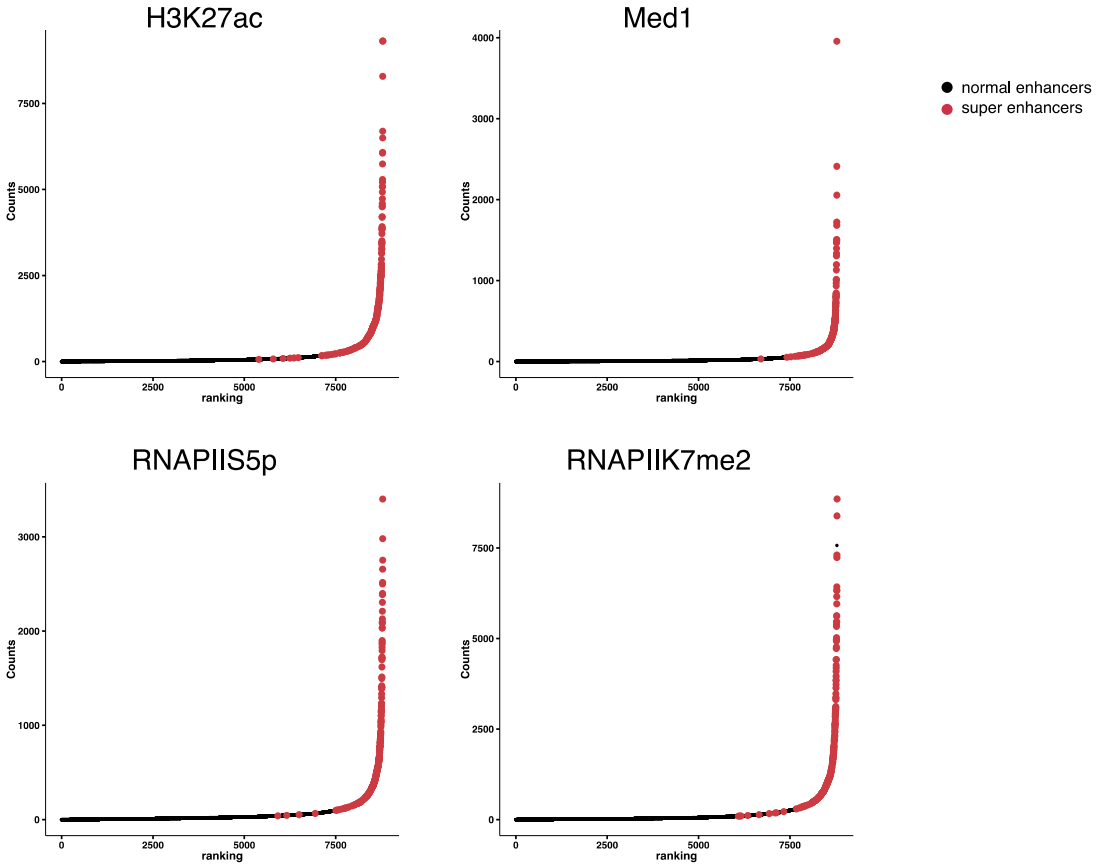


Fig.4.14: RNAPII performs good distinguishing normal and super enhancers. a) Ranking analysis over all Whyte enhancers for SE factors and RNAPII modifications. Enhancers were ranked from less to more enriched for each modification. Zoom in shows the enrichment turning point used to distinguish normal enhancers and super enhancers. b) Ranking plot of enrichment for Med1, H3K27ac, RNAPIIS5p, RNAPIIK7me2. Enrichment was calculated per Whyte enhancer region and regions were ordered for least to most enriched, as in the original Whyte paper. In black: enhancers; in red: super enhancers defined in Whyte et al. 2013.

4.5 Discussion

RNA polymerase II is responsible for the transcription of all protein-coding genes, lincRNAs, small RNAs and enhancer RNAs. Although its presence at regulatory regions had been observed before, a deep investigation of RNAPII activation states and the relationship with enhancer activation state was not previously conducted.

4.5.1 RNAPII exists in different states of activation at enhancers

In the current chapter, I investigated the state of RNAPII at enhancer regions previously defined based on Nanog, Sox2, and Oct4 co-occupancy (Whyte et al., 2013), and I focused on active enhancers. I only analysed extragenic regions to avoid confounding effects of detecting initiating or elongating RNAPII at promoter, coding and termination regions of genes. Interestingly, more than 50% of enhancer regions identified using histone marks, TFs or chromatin factors map to intragenic regions (data not shown), irrespectively of the list of enhancers analysed (Arner et al., 2015; Chen et al., 2012; Creyghton et al., 2010; Zentner et al., 2011).

To achieve highest confidence on the RNAPII presence at enhancer regions, I took into account that the RNAPII occupancy outside coding regions is low (De Santa et al., 2010). To this end I developed a classification strategy that considers the noise occupancy across the genome, with random permutated region and calculation of a 5% False Positive cut-off. Combined with the use of BCP to detect positive windows (Xing et al., 2012), the classification also considers local noise intrinsic of ChIP-seq experiments. BCP models the enrichment of the feature of interest over the enrichment of a matched control immunoprecipitation (mock immunoprecipitations, in the present study). The chosen approach aims to give high confidence on the classification of RNAPII presence, at the expense of the number of positive regions.

RNAPII classification at extragenic enhancers showed that RNAPII is present in different states, from initiating RNAPIIS5p to a fully elongating one, with RNAPIIS7p and RNAPIIS2p. These modifications encompass the states described at genes. Interestingly, RNAPII at enhancers can also be found in an unphosphorylated form, RNAPIIS2u, which is uncommon at genes (Brookes et al., 2012).

The same analyses presented here were also conducted using other published enhancer lists (Arner et al., 2015; Chen et al., 2012; Creyghton et al., 2010; Zentner et al., 2011), reaching similar conclusions about RNAPII presence and activation at enhancers regardless of the initial approach used to define enhancers. Most interestingly, the nature of features used to define each enhancer list influences the proportion of RNAPII-bound enhancers recovered. For example, the

Arner list (Arner et al., 2015) defined enhancers as regions of the genome marked by CAGE tags outside gene promoters, such that by definition the Arner's enhancers are transcribed and the transcripts produced are at least capped. As eRNA transcription is linked with active enhancers (Kim et al., 2010;Kaikkonen et al., 2013), most of the Arner enhancers are bound by RNAPII and in an elongating form. In contrast, the Whyte enhancer list is composed of enhancers bound by TFs: the enhancers contained in this list are also defined as active. While the Whyte list does not include poised enhancers, they nevertheless present different levels of activation based on RNAPII states or chromatin marks. Interestingly, SEs are associated with active RNAPII state, which suggests a role of RNAPII state in enhancer activation. The fact that Whyte enhancers do not show enrichment of repressive marks is intriguing. Based on results from the previous chapter, it appears that TF binding differs at active and Polycomb repressed enhancers: when Nanog, Sox2, and Oct4 are found together, such in the case of Whyte enhancers, enhancers are in active state, while if only one or two are bound, enhancers are more likely to be repressed. It would be interesting to understand if other factors are involved in this difference and if they recruit chromatin remodellers. Finally, the analysis of the Chen enhancer list (Chen et al., 2012) showed that enrichment of H3K27me3, a mark of Polycomb repression, occurs at RNAPIIS5p classes of enhancers (data not shown), in line to what we know at genes (Brookes et al., 2012; Ferrai et al., 2017; Stock et al., 2007; Tee et al., 2014). This topic will be further investigated in the next chapters.

4.5.2 RNAPII is associated with enhancers in different activation states

The enhancers in the Whyte list are identified by the co-binding of Nanog, Sox2 and Oct4. By definition these regions are considered to be all similar and have an active state, but through the analyses of RNAPII, histone modifications and transcription factors occupancy, I find that not all enhancers in the Whyte list are the same, as they have different properties and features. Noticeably, RNAPII-bound Whyte enhancers are longer than unbound ones and have broader RNAPII peaks. Interestingly, Nanog, Sox2 and Oct4 do not have a strong differential enrichment at these regions, while other TFs such as E2f1 or Klf4 preferentially bind enhancers with active states of RNAPII and Smad1 with primed RNAPIIS5p. These differences would indicate that among the Whyte enhancer regions, some are more activated than the others. Based on the current literature, enhancer regions enriched for factors such as P300 and H3K27ac would be more likely to be the most active (Creyghton et al., 2010), which suggests that active RNAPII states mark the most active enhancers. In mESCs, enhancers bound by RNAPII are associated with genes related to stem cell maintenance and negative regulators of differentiation, while

enhancer regions not bound by RNAPII are associated with gene terms more related with early differentiation and exit from pluripotency. Possibly, RNAPII-bound enhancers keep the mESC state, while enhancers not bound by RNAPII are less active, or fluctuating within the ESC population, ready to differentiate. This observation suggests that enhancers in an active state can be further dissected in more subtle and possibly biologically meaningful subgroups.

To understand whether different states of RNAPII can more finely dissect enhancer activation state, I analysed different features. H3K27ac is gradually enriched in concordance with RNAPII activation state and increased RNAPII occupancy levels. In 2010, Creyghton and colleagues (Creyghton et al., 2010) showed that enhancers are increasingly active the more H3K27ac marks is enriched, and in 2013 Whyte *et al.* 2013 (Whyte et al., 2013) showed that stretches of enhancers highly active have a higher enrichment for marks such as H3K27ac and Med1, with similar TFs binding. These results suggest that RNAPII state is indicative of enhancer activation state. The more RNAPII is found in an active state the more regulatory regions are occupied by marks associated with their activity. Active enhancers can therefore be divided in more and less active, according to RNAPII occupancy. Interestingly, the results of this chapter show differences between enhancers bound with RNAPII configurations typical of productive transcription at active genes (RNAPIIS2u-S5p, RNAPIIS2u-S5p-S7p, RNAPIIS2u-S5p-S7p-S2p) and also with regions associated with paused (RNAPIIS2u) or Polycomb-poised states (RNAPIIS5p alone). Concordant with this observation, RNAPIIS2u and RNAPIIS5p classes of enhancers show an intermediate enrichment for active chromatin marks, which hints at different transcriptional states and potentially enhancer activity, which will be partially analysed in the following chapters.

Activity at enhancers is always difficult to define: it could be the strength or the stability of target gene enhanced expression. It will be interesting to analyse whether different forms of RNAPII contribute to these effects with different mechanisms or to different degrees.

The finding that RNAPII is present at different state of enhancers that mirror enhancer activation state offers a new layer in understanding enhancer activation. TF binding at enhancers is the first step, however it is not indicative of the state of activity of an enhancer: chromatin marks, and now RNAPII state, can give a deeper of understanding of enhancer activation state.

4.6 Figures Appendix

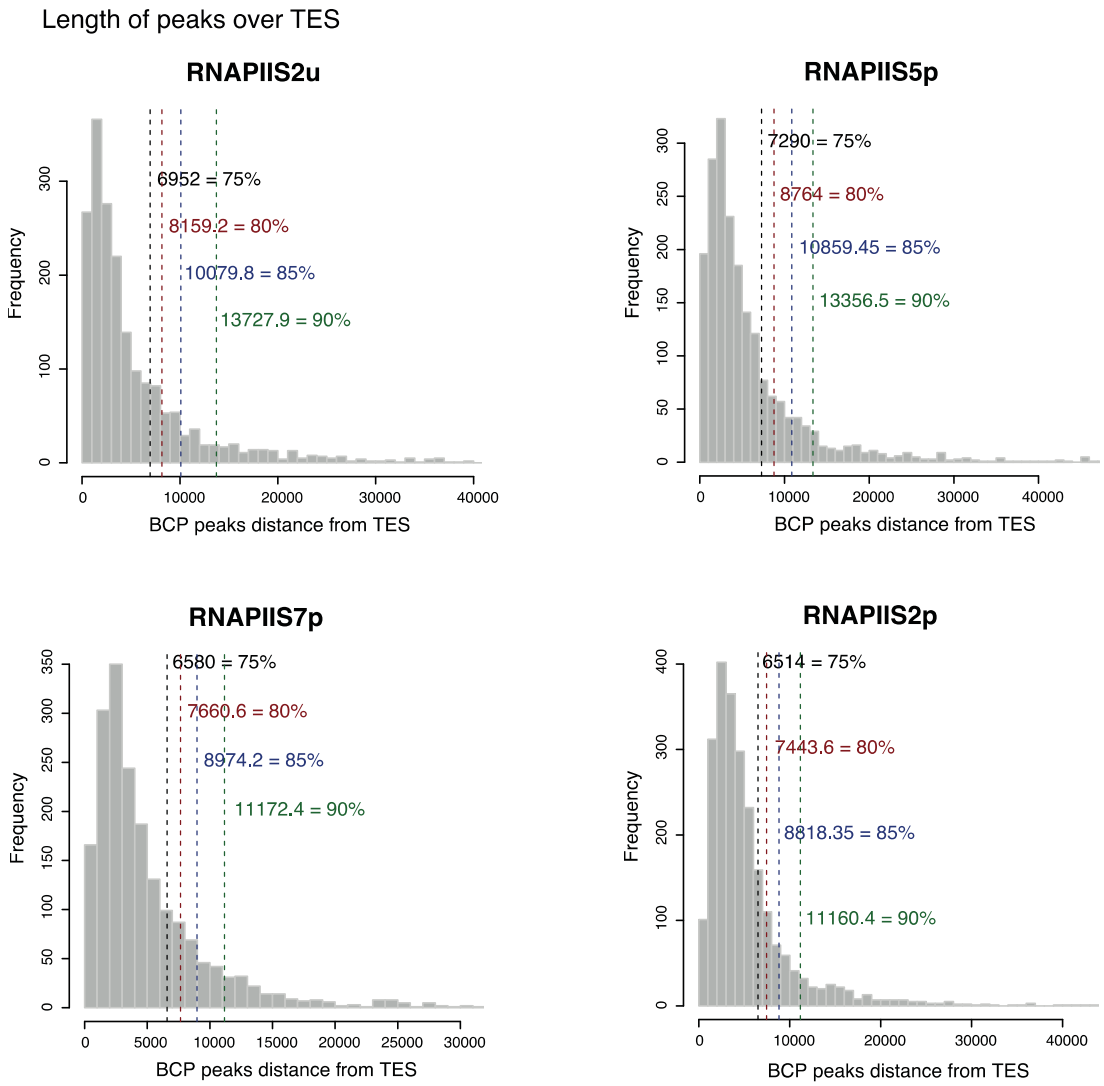


Fig 4.A1: Distribution of RNAPII peaks length overhanging TESs in mESC. Coloured numbers represent the length below which 75%, 80%, 85%, 90% of all RNAPII peaks are found, respectively. Datasets of RNAPII modifications from Brookes et al. 2012.

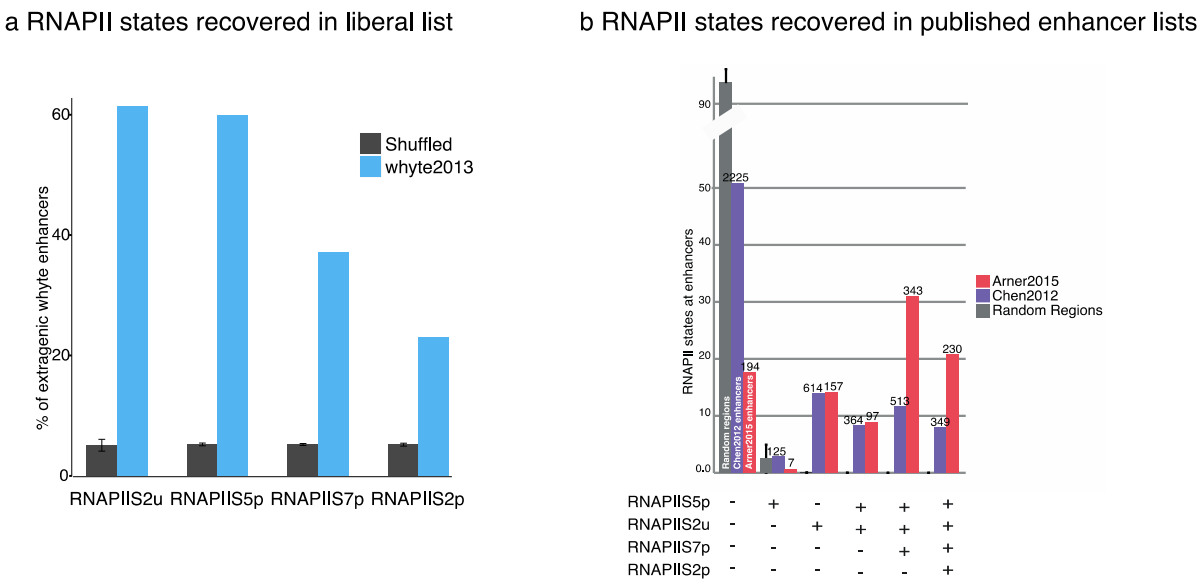


Fig 4.A2: Classification of extragenic enhancers. a) Percentage of each RNAPII modification at Whyte extragenic enhancers classified with the liberal approach. On the x-axis: RNAPII modification. On the y-axis: percentage of extragenic Whyte enhancers positive for that modification. Blue bars represent Whyte enhancers. Grey bars represent extragenic permutated regions. Error bars are computed on randomly permuted regions classification and indicate standard deviation. b) Classification of enhancers from Chen et al. 2012 and Arner et al. 2015 ordered by RNAPII activation. On the x-axis: state of RNAPII. On the y-axis: number of regions. Purple bars represent extragenic Chen 2012 extragenic. Light red bar represent extragenic Arner 2015 extragenic enhancers. Grey bars represent extragenic permutated regions (30 times). Error bars are computed on randomly permuted regions classification and indicate standard deviation.

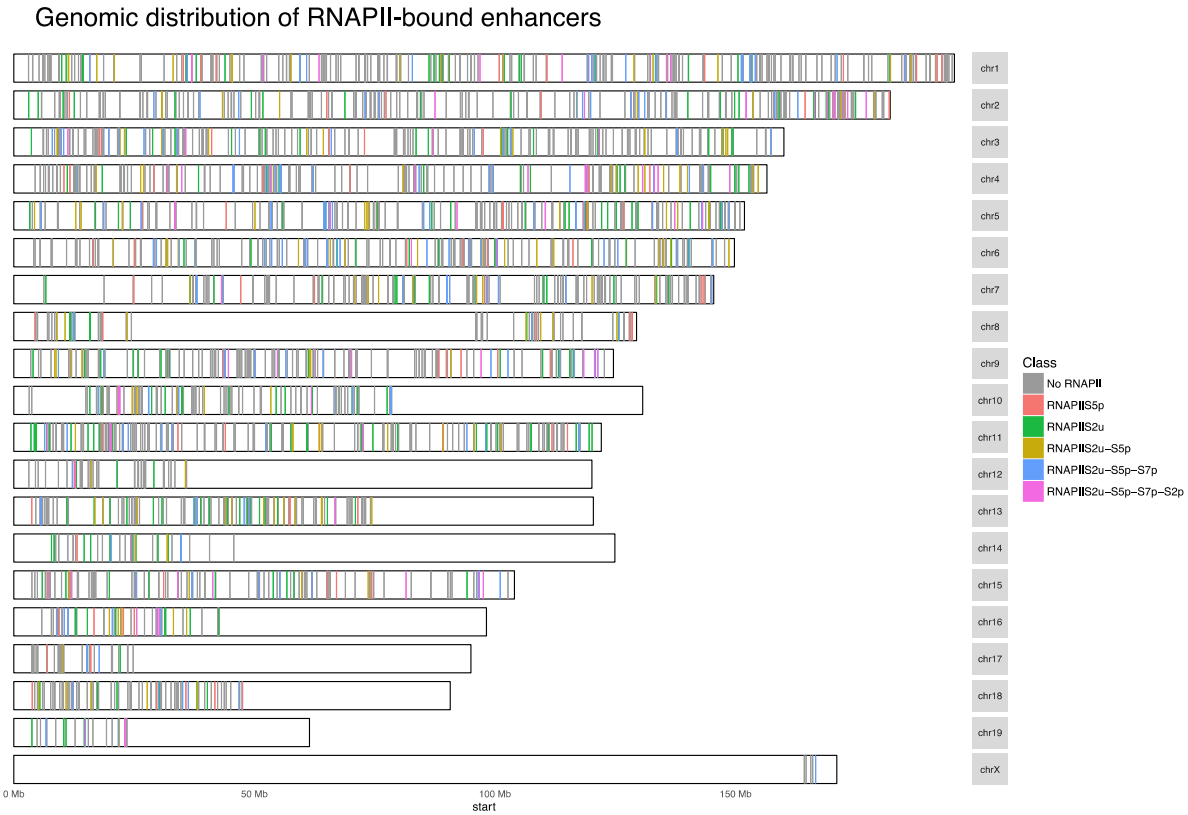


Fig 4.A3: Genomic distribution of extragenic Whyte enhancers. Distribution of RNAPII-bound and not bound extragenic Whyte enhancers in different chromosomes. White spaces are gene dense regions with no extragenic enhancers, based on our classification. Each vertical line represent an enhancer location, color-coded for the enhancer class with the maximum divergent colours to enhance readability

5. Extragenic RNAPII states identify enhancers in different activation states

5.1 Introduction

In the previous chapter I studied the presence and state of activation of RNAPII at published lists of enhancers and found that RNAPII occupies known extragenic enhancers with different activation states. In the current chapter, I will identify extragenic RNAPII peaks independently of previous enhancer lists, and investigate their association with enhancer features, and candidate roles as enhancers.

5.1.1 Approaches to identify regulatory regions

One major challenge in the enhancer field remains the identification of putative regulatory regions. Numerous approaches and advances were made in recent years, and each approach identifies and categorises different regions (Chapter 3 for more details). On the one hand, TF binding is a reliable and powerful approach to identify candidate enhancer regions, with the requirement to produce or have access to numerous ChIP-seq datasets for several transcription factors important for each cell type and upon response to stimuli, including the availability of antibodies. However, TF binding fails to recapitulate the different states of activation of enhancers (Fig 4.11, see also (Hnisz et al., 2013; Whyte et al., 2013)), which in turn can be distinguished through the levels of occupancy of other chromatin marks, such as H3K27ac (Creyghton et al., 2010). Chromatin marks, on the other hand, mark very broad regions and can be unspecific to point the precise enhancer location (Calo and Wysocka, 2013).

I was interested in exploring the possibility of defining enhancers and their activation state using a minimal number of datasets, which themselves in addition directly inform of the state of gene promoters and coding regions. Towards defining novel ways of identifying candidate enhancer regions, I explored the use of RNAPII and Polycomb datasets to dissect the state of extragenic regulatory regions.

5.1.2 Transcription regulation at genes

RNAPII activation states are highly regulated at genes. RNAPII is post-translationally modified during the transcription cycle and these modifications act as recruiters for chromatin remodelers and RNA maturation machineries which act co-transcriptionally (Brookes and Pombo, 2009; Egloff and Murphy, 2008). Therefore, the state of RNAPII and its regulation influence the chromatin environment and the RNA products from a given gene.

Interestingly, RNAPII phosphorylation on Ser5 at Polycomb-repressed genes is carried out by Erk2 kinase, instead of Cdk7 (Tee et al., 2014). Polycomb repressed genes are also regulated by Utf1, that on the one hand prevent the excessive spreading of PRC2 mark H3K27me3 and on the other recruits the de-capping machinery that regulates unnecessary RNA production from repressed genes (Jia et al., 2012) (Scheme of regulation at genes in Fig 5.1). It is unknown if these proteins are active at extragenic poised regions.

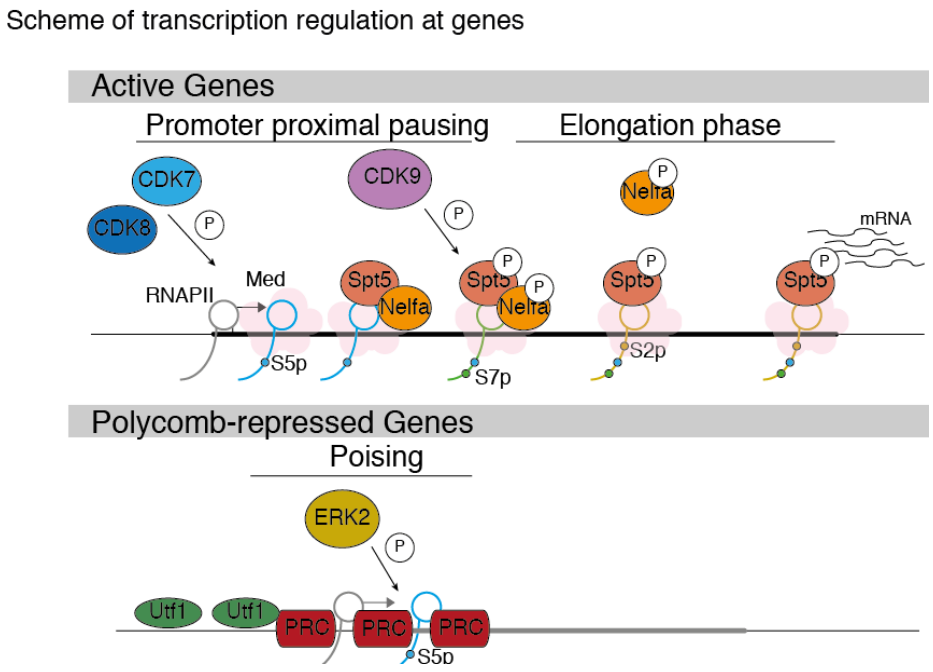


Fig 5.1: Regulation of transcription at coding genes. Scheme representing regulation of transcription at active genes (on top), and Polycomb Repressed genes (PRC) on bottom.

Another interesting phenomenon is the transcription of short and unstable RNAs upstream of active genes (Almada et al., 2013; Core et al., 2014; Preker et al., 2011). Even if it is still debated whether all active promoters are bidirectionally transcribed or only a subset, some interesting features were described. Transcripts produced upstream of promoters (PROMPTs) are usually very unstable and sensitive to exosome degradation (Preker et al., 2008), not polyadenylated and not spliced (Fig 5.2). These similarities of eRNAs and PROMPTs are currently explored in the field.

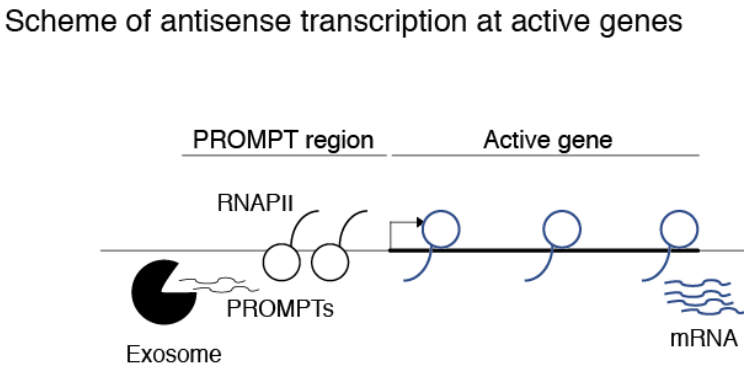


Fig 5.2: Antisense transcription at active genes. Scheme representing sense and antisense transcription at active promoters. Antisense transcripts, here defined as PROMPTs, are exosome sensitive.

It is unknown to what degree transcription of extragenic RNAPII (RNAPII not transcribing genic regions) is regulated. Previous studies have shown that eRNA levels decrease after Cdk9 inhibition through Flavopiridol, without effect on enhancer-promoter looping (Hah et al., 2013), and that eRNAs are sensitive to Exosome degradation (Andersson et al., 2014b).

5.2 Aim of the chapter

In the current chapter, I have defined extragenic RNAPII regions independently of previous enhancer classifications and analysed their chromatin and transcriptional state to understand to what extent they help to identify extragenic enhancers. The results of the current chapter show that extragenic RNAPII recognises new candidate regulatory regions. These regions are regulated at the transcriptional and the post-transcriptional levels, responding to similar regulators as RNAPII at genes and at other described enhancers. RNAPII is therefore a powerful approach to identify putative enhancers in a cell type of interest, and without further work to inform of enhancer activation states.

5.3 Contribution disclosure

Elena Torlai Triglia mapped the RNAPII and H3K27me3 datasets and calculated their peaks, and re-mapped Brd4, CTCF, Utf1 and Erk2 ChIP-seq datasets. Some of the datasets in the current chapter are unpublished, as follows: Nascent RNA-ChIP-seq in control and Flavopiridol-treated mESCs, Total RNA after Exosome knockdown. These datasets were produced by Kelly J. Morris and Robert A. Beagrie. 46C Nascent RNA-ChIP untreated was mapped by Robert A. Beagrie. The remaining datasets were mapped by me. Exosome Dis3 ChIP-seq dataset was produced by Ana Miguel Fernades and analysed by Elena Torlai Triglia. Total RNA-seq datasets were produced by Ana Miguel Fernandes and Carmelo Ferrai and mapped by me.

5.4 Results

5.4.1 Strategy to define extragenic RNAPII regions

Inspection of ChIP-seq tracks reveals the presence of RNAPII at extragenic regions not previously classified as enhancers (Fig. 5.3).

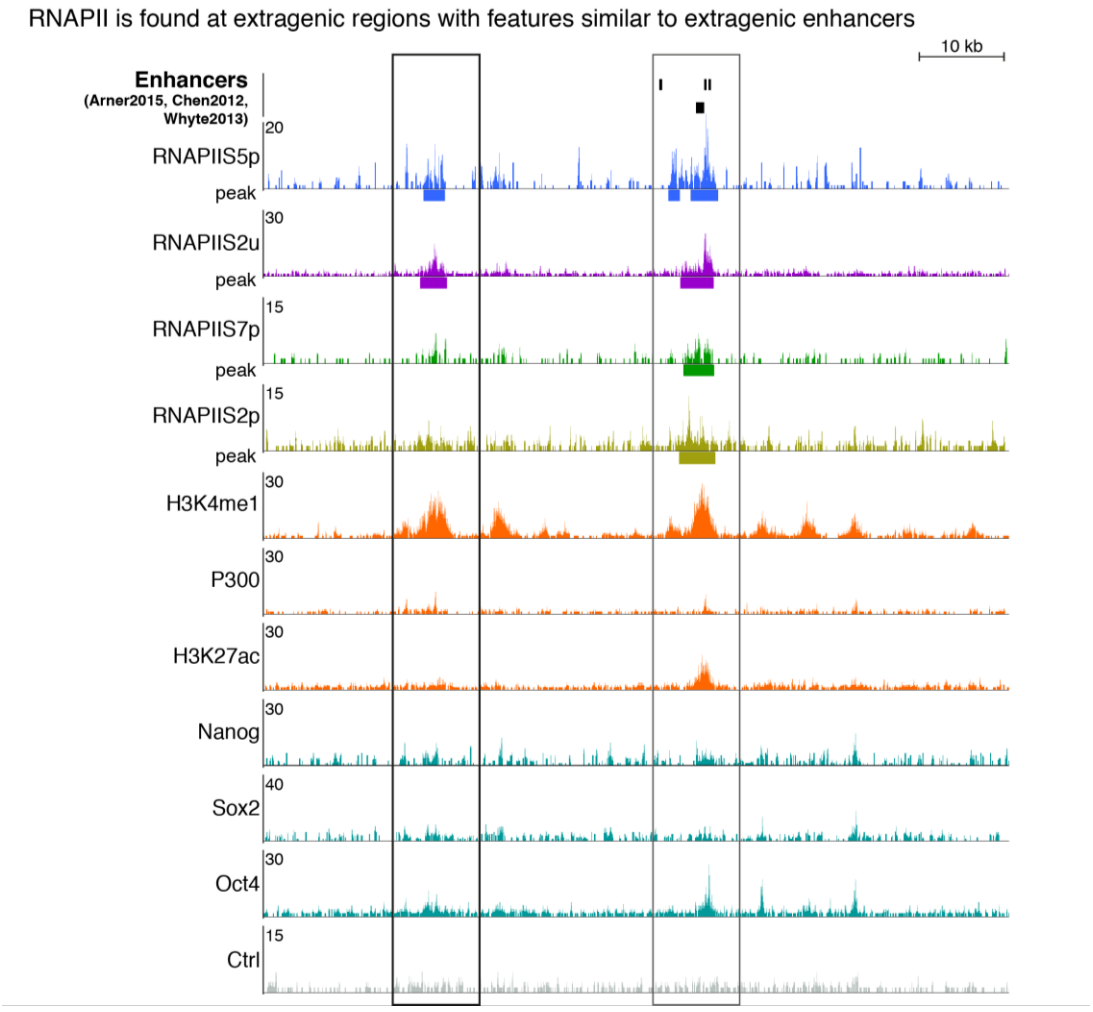


Fig 5.3: Extragenic RNAPII mark putative regulatory regions missed by other approaches. UCSC screenshot of extragenic regions. Black rectangles represent enhancers identified in Whyte et al. 2013 1, Arner et al. 2015 29, Chen et al. 2012 19. Tracks always show the 0 to the maximum high indicated, with a smoothing window of 2. Coloured boxes under RNAPII tracks represent BCP peaks for that dataset. Empty black box highlight extragenic regions marked by RNAPII and enhancer marks. Image generated with UCSC genome browser.

To define extragenic RNAPII regions, I used an approach similar to the one used in Chapter 4 and considered as extragenic all the RNAPII peaks that do not overlap with a gene or 2kb upstream the TSS annotated in the RefSeq list (O'Leary et al., 2016) or in the UCSC gene list (Casper et al., 2018) and any RNAPII peaks that do not contiguously extend beyond annotated coding region (Fig. 5.4a). With this approach, some intragenic regions which could potentially function as enhancers are not considered, but being conservative enables to explore RNAPII as an enhancer marker minimizing the confounding effects of non-annotated alternative

transcription start sites upstream of gene promoters. All regions overlapping with the ENCODE blacklist were also excluded from the present analysis, to remove genomic regions with "anomalous, unstructured, high signal/read counts in next gen sequencing experiments independent of cell line and type of experiment" (Consortium, 2012).

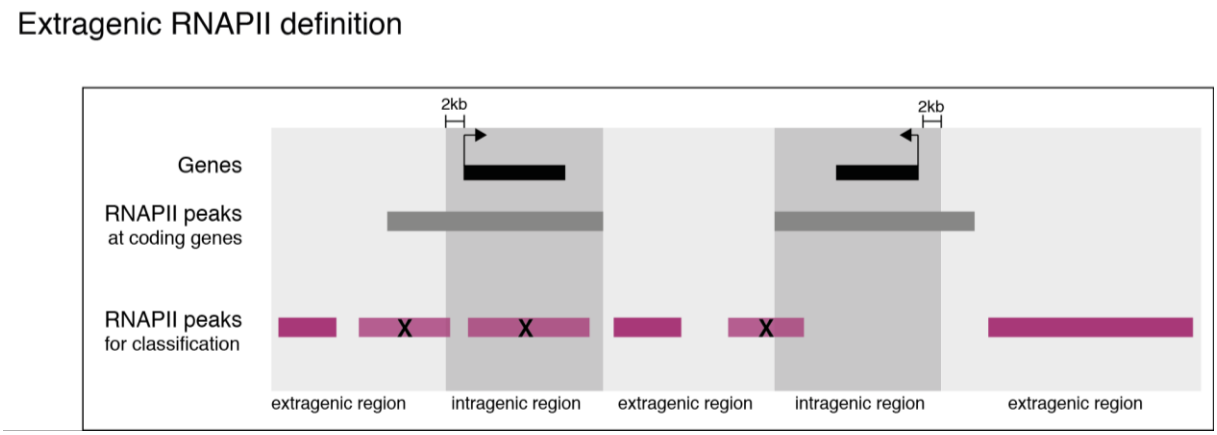


Fig 5.4: Extragenic RNAPII peaks definition. Scheme of extragenic definition of RNAPII peaks in the current chapter. Black rectangles represent genes, grey rectangles represent RNAPII peaks overlapping coding genes. Purple rectangles represent RNAPII peaks to be divided in extragenic and intragenic. X represents a intragenic peak and therefore not analysed in the following paragraphs. Transparent light grey boxes define extragenic regions; transparent dark grey boxes define intragenic regions.

5.4.2 Strategy to classify extragenic RNAPII regions

To identify RNAPII extragenic regions and to investigate to what extent they coincide with previously described enhancer regions, I first choose the RNAPII datasets that could be most informative for this work. First, I compared different RNAPII modifications to understand whether they co-localise at extragenic regions and to analyse their chromatin features occupancy. RNAPII dataset from Brookes *et al.* 2012 (Brookes et al., 2012), namely RNAPIIS5p, RNAPIIS7p, RNAPIIS2u (RNAPIIS2 unpshopshorylated), were classified at extragenic or intragenic (Fig. 5.5a), as described previously in chapter 4 (4.4.3). RNAPIIS2p was not considered in these analyses because of the low number of extragenic RNAPIIS2p peaks (737) and the results from the previous analysis in Chapter 4 (4.4.7) that showed the co-occurrence of RNAPIIS2p with RNAPIIS7p; RNAPIIS2p therefore would not be able to discover new regions. Next, the enrichment in each RNAPII modification was determined at extragenic peaks and random (shuffled) regions with similar properties as the identified peaks. Finally, a 5% cut-off was calculated based on the random region enrichment. This procedure identified 5319 RNAPIIS2u extragenic regions (96% were above threshold), 3699 RNAPIIS5p regions (84%) and 2273 RNAPIIS7p regions (96%) (Fig 5.5b).

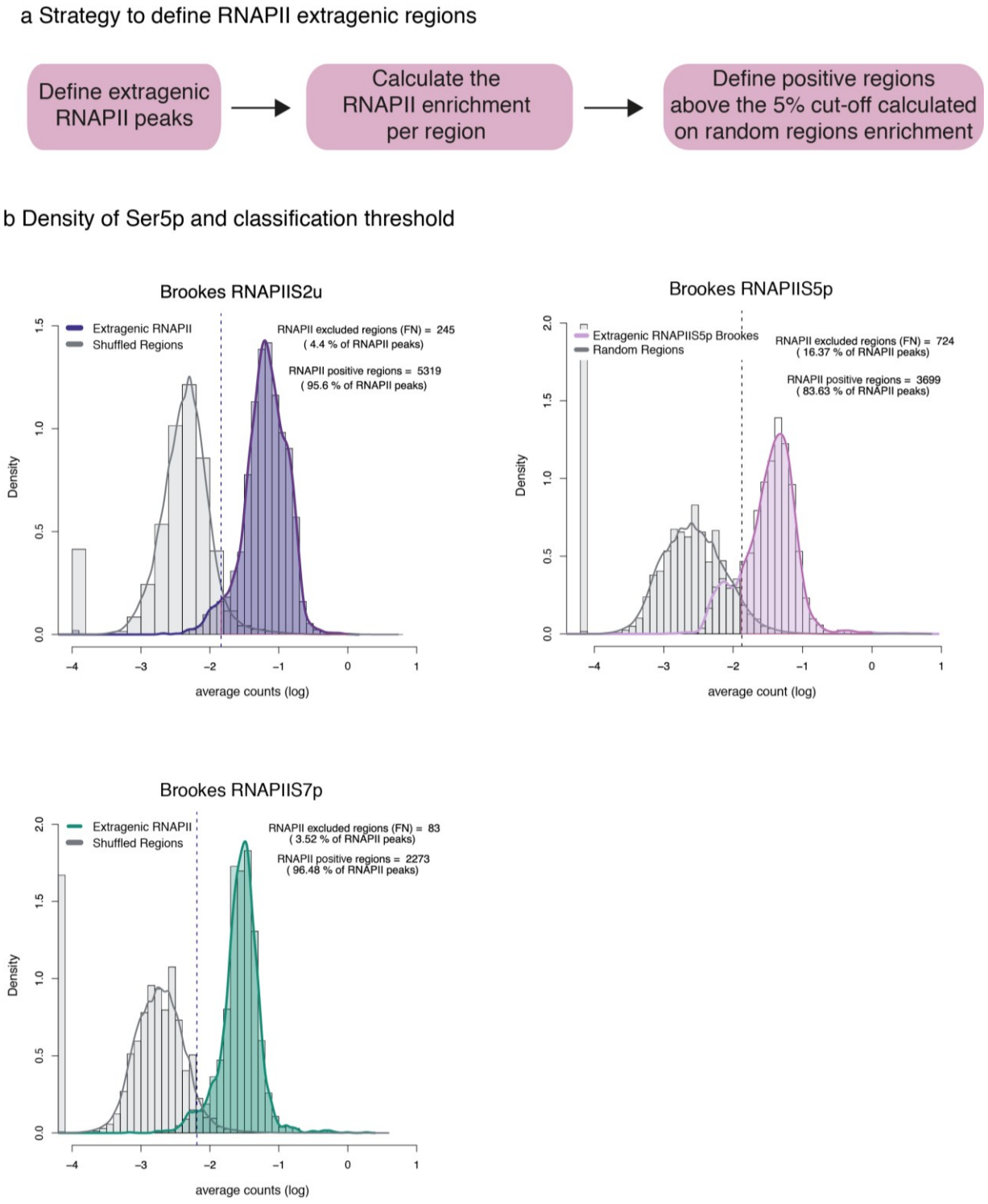


Fig 5.5: Definition of extragenic RNAPII Brookes datasets. a) Strategy to define RNAPII extragenic regions. b) Classification of RNAPIIS2u, RNAPIIS5p, RNAPIIS7p extragenic regions. Coloured thick lines represent enrichment of RNAPII modification at RNAPII extragenic regions. Grey lines represent enrichment at random extragenic regions. Dotted vertical line represent 5% FP cut-off.

In conclusion, I generated three extragenic RNAPII lists of regions with different RNAPII modifications to then investigate whether they identify the same regions and which properties these regions have.

5.4.3 Comparison of RNAPII modifications at extragenic regions

To understand how the different extragenic RNAPII regions relate with each other. I first analysed their co-localisation in the genome. Most RNAPII extragenic regions are marked by both RNAPIIS2u and/or RNAPIIS5p (4335 regions), with 1814 regions are found with the three RNAPII datasets. Only 130 regions are identified by RNAPIIS7p alone (Fig 5.6a), in line with the results of the previous chapter and with RNAPII modification at coding genes, where RNAPIIS7p occurs downstream of RNAPIIS5p on polymerases also marked by RNAPIIS2u. RNAPIIS5p and RNAPIIS2u alone also identify unique regions: 2152 (40%) and 728 (20%), respectively.

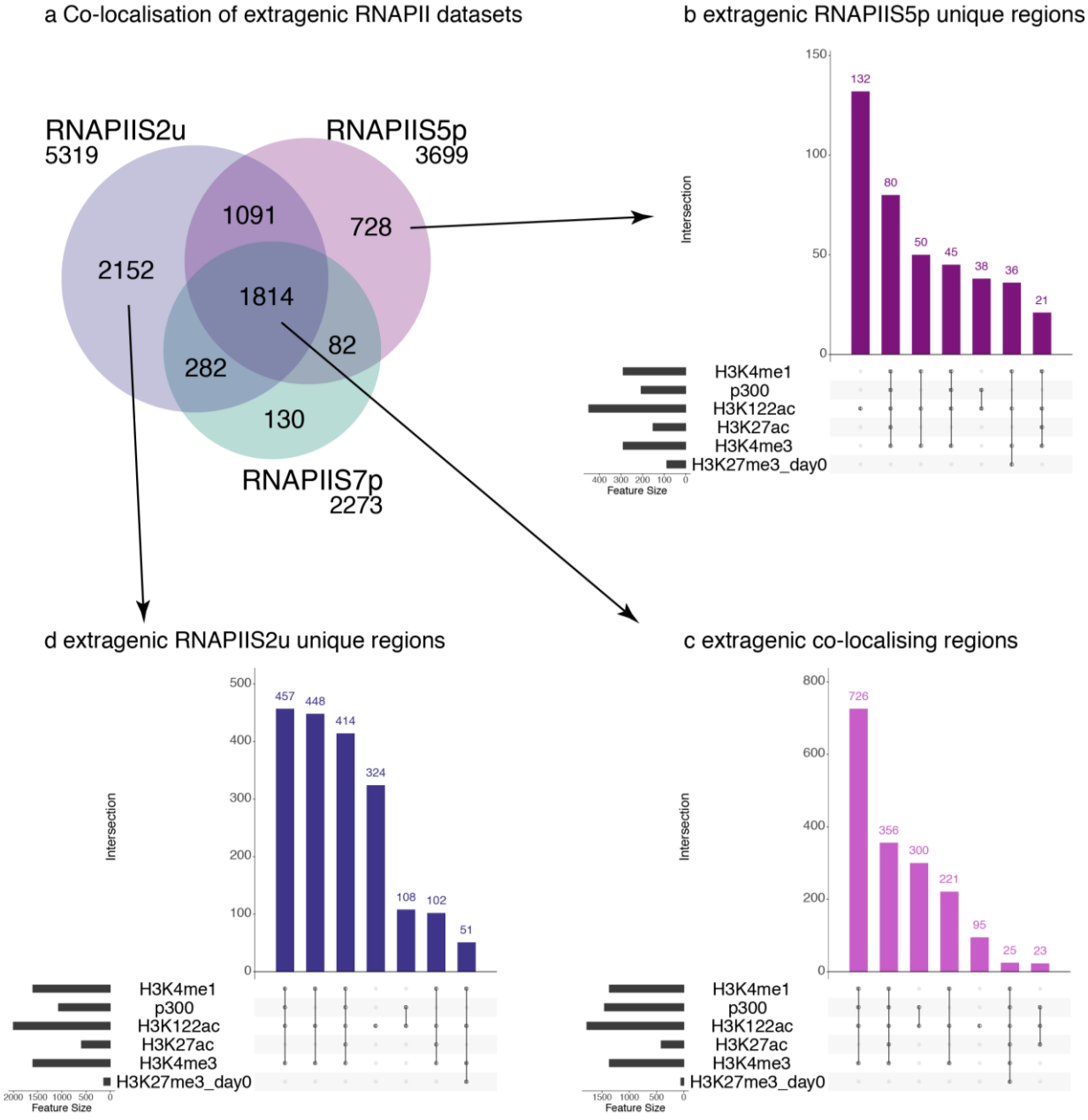


Fig 5.6: Extragenic RNAPII modifications identify similar regions in the genome. a) Venn diagram showing the co-localisation of extragenic RNAPII peaks for three different datasets from Brookes et al: RNAPIIS2u, RNAPIIS5p, RNAPIIS7p. The numbers indicate the number of region per group. b) Occupancy of histone modifications and co-factors at regions identified solely by RNAPIIS5p (right on top, in dark pink) c) identified by RNAPIIS2u, RNAPIIS5p and RNAPIIS7p (right center, in pink) d) identified solely by RNAPIIS2u datasets (left bottom, in purple).

To explore the candidate enhancer roles and regulation of the extragenic RNAPII regions, I measured their overlap with H3K4me1, p300, H3K122ac, H3K27ac, H3K4me3, and H3K27me3. Regions marked by all three RNAPII modifications are most often co-occupied by all tested enhancer marks, and only a small number is bound by H3K27me3, as expected from the lack of co-association between Polycomb marks and elongating RNAPII at active genes (Brookes et al., 2012). The regions uniquely found with RNAPIIS2u and RNAPIIS7p differ slightly in the proportion of occupancy of important enhancer features. RNAPIIS2u regions are mainly occupied by H3K4me1, in combination with marks of non-canonical enhancers (H3K122ac) and H3K4me3, a mark of open promoters. RNAPIIS5p unique regions are more occupied by non-canonical enhance marks and proportionally more occupied by H3K27me3 (Fig 5.6b, c, d).

To investigate the potential of extragenic RNAPII occupancy to define enhancers, I compared how many of the RNAPII peaks identified overlapped with previously identified enhancers using the Chen *et al.* 2012 (Chen et al., 2012), Whyte *et al.* 2013 (Whyte et al., 2013), Cruz Molina *et al.* 2017 (Cruz-Molina et al., 2017) lists. I found that 30%, 22% ,and 22% of the S2u, S5p and S7p peaks coincided with an enhancer region previously identified (data not shown), confirming their partial redundancy with alternative criteria to detect candidate enhancers, but also the potential of using RNAPII peaks to discover new candidate regions and their activation states. This topic will be explored in more detail in the following chapter (chapter 6).

From these analyses, it was clear that both RNAPIIS2u and RNAPIIS5p identify candidate extragenic enhancers. The following analyses presented in this chapter focus on studying the new candidate enhancers only on RNAPIIS5p and not RNAPIIS2u for two reasons: first, RNAPIIS5p is present at repressed genes together with Polycomb and a poised enhancers, while RNAPIIS2u would miss these regions; second, genome-wide mapping of RNAPIIS5p but not RNAPIIS2u is available in a time-course neuronal differentiation (Ferrai et al., 2017) which allows not only the exploration of extragenic RNAPIIS5p regions in ESCs, but also the opportunity to follow dynamics of candidate RNAPII-marked enhancers during differentiation. A brief inspection of the 2152 RNAPIIS2u extragenic regions show that 60% do not overlap with known enhancers (Whyte *et al.* 2013 (Whyte et al., 2013), Cruz Molina *et al.* 2017 (Cruz-Molina et al., 2017)), and that ~50% coincide with TF binding sites (data not show); these regions will be explored in future analyses.

5.4.4 Classification of extragenic RNAPIIS5p regions in the mESC 46C clone

Ferrai *et al.* 2017 (Ferrai et al., 2017) have produced ChIP-seq datasets for RNAPII-S5p, S7p, S2p, H3K27me3 and mRNA-seq during a differentiation time line that starts in ESCs using clone 46C (day 0), covers the early exit from pluripotency (days 1 and 3) and ends with immature and mature dopaminergic neurons (days 16 and 30, respectively). In this chapter, I study the ESC datasets (Table 5.1), and in the following chapter I will explore datasets for all time points. Another publication from the lab, Fraser *et al.* 2015 (Fraser et al., 2015), describes Hi-C and CAGE data produced from ESCs (day 0) and immature neurons (day 16). Other important unpublished data was available in the laboratory from the same ESC clone and timeline, namely nascent and total RNA. Therefore, to leverage on these resources which are matched for the same ESC clone and differentiation timeline, I decided at this stage to transfer my analyses of RNAPII extragenic enhancers from the Brookes datasets analysed (which included more RNAPII marks) to the published datasets from Ferrai *et al.* 2017 (Ferrai et al., 2017). To this end, I repeated the identification of extragenic RNAPII-S5p peaks using published RNAPIIS5p produced in mESC 46C datasets as previously described for Brookes RNAPII datasets.

Table 5.1: RNAPII datasets from Ferrai et al. 2017 Table describing RNAPII dataset used in the following analysis. Datasets mapped and processed by Elena Torlai Triglia.

Dataset	Number of peaks	mESC clone	Publication	GEO
RNAPIIS5p	24010	46C	Ferrai *et al.*, 2017	GSM2474111
RNAPIIS7p	17485	46C	Ferrai *et al.*, 2017	GSM2474112
RNAPIIS2p	8139	46C	Ferrai *et al.*, unpublished	Ferrai *et al.*, unpublished
H3K27me3	6677	46C	Ferrai *et al.*, 2017	GSM2474113

In brief, I divided previously calculated RNAPIIS5p peaks by Elena Torlai Triglia in extragenic and intragenic and excluded regions overlapping with the ENCODE blacklist. I then calculated the enrichment for RNAPIIS5p from the ESC (Day 0) dataset and classified them as positive if they were enriched over a 5% false positive cut-off calculated on matched extragenic random regions (Fig 5.7a). More than 83% (4432) of RNAPIIS5p Day 0 peaks were above the enrichment cut-off, which is comparable with the RNAPIIS5p Brookes dataset.

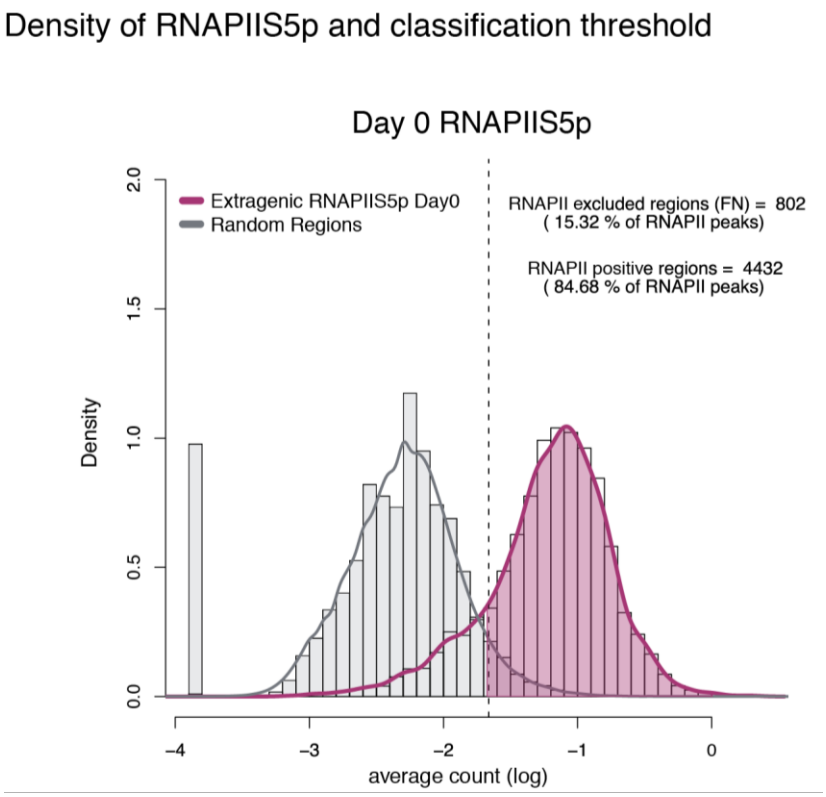


Fig 5.7: Extragenic RNAPIIS5p day0 definition. a) Classification of RNAPIIS5p extragenic regions. Purple thick lines represents enrichment of RNAPIIS5p at RNAPIIS5p extragenic regions. Grey lines represent enrichment at random extragenic regions. Dotted vertical line represent 5% FP cut-off.

5.4.6 Comparison of RNAPIIS5p extragenic regions in two ESC lines

To compare extragenic regions occupied by RNAPIIS5p in ESC clones OS25 (Brookes *et al.* 2012 (Brookes et al., 2012)) and 46C (Ferrai *et al.* 2017 (Ferrai et al., 2017)), I first analysed their length distribution, also against published enhancer previously analysed. The two RNAPIIS5p datasets were produced with the same antibody clone against RNAPII-S5p (clone CTD-4H8), using the same chromatin preparation and similar ChIP and sequencing strategies. Importantly, the two ESC lines are grown in different conditions: OS25 cells are grown in serum+LIF conditions under selection for Oct4 expression, whereas 46C cells are grown in serum free conditions. Both conditions preserve the repression of Polycomb-repressed developmental genes, but show noticeable differences in the expression of some signalling genes (Elena Torlai Triglia, personal communication), as expected from the different growth conditions.

Extragenic RNAPIIS5p regions from Brookes and Ferrai datasets had comparable lengths, and both tended to be longer than the published Whyte enhancers defined with TF binding (Fig 5.8a), irrespectively of whether the full extragenic Whyte list was considered or only the extragenic Whyte enhancers occupied by RNAPIIS5p. Comparisons between the two RNAPIIS5p datasets

showed that ~2500 extragenic RNAPIIS5p regions are common to the two cell lines, with ~4000 RNAPII extragenic regions identified exclusively in each dataset (Fig 5.8b). These observations highlight the importance of matching datasets for enhancer regions on the same cell line and growth conditions, as it is expected from comparisons between different cell types that enhancer marking varies more than expressed gene cohorts (Anderson and Hill, 2014).

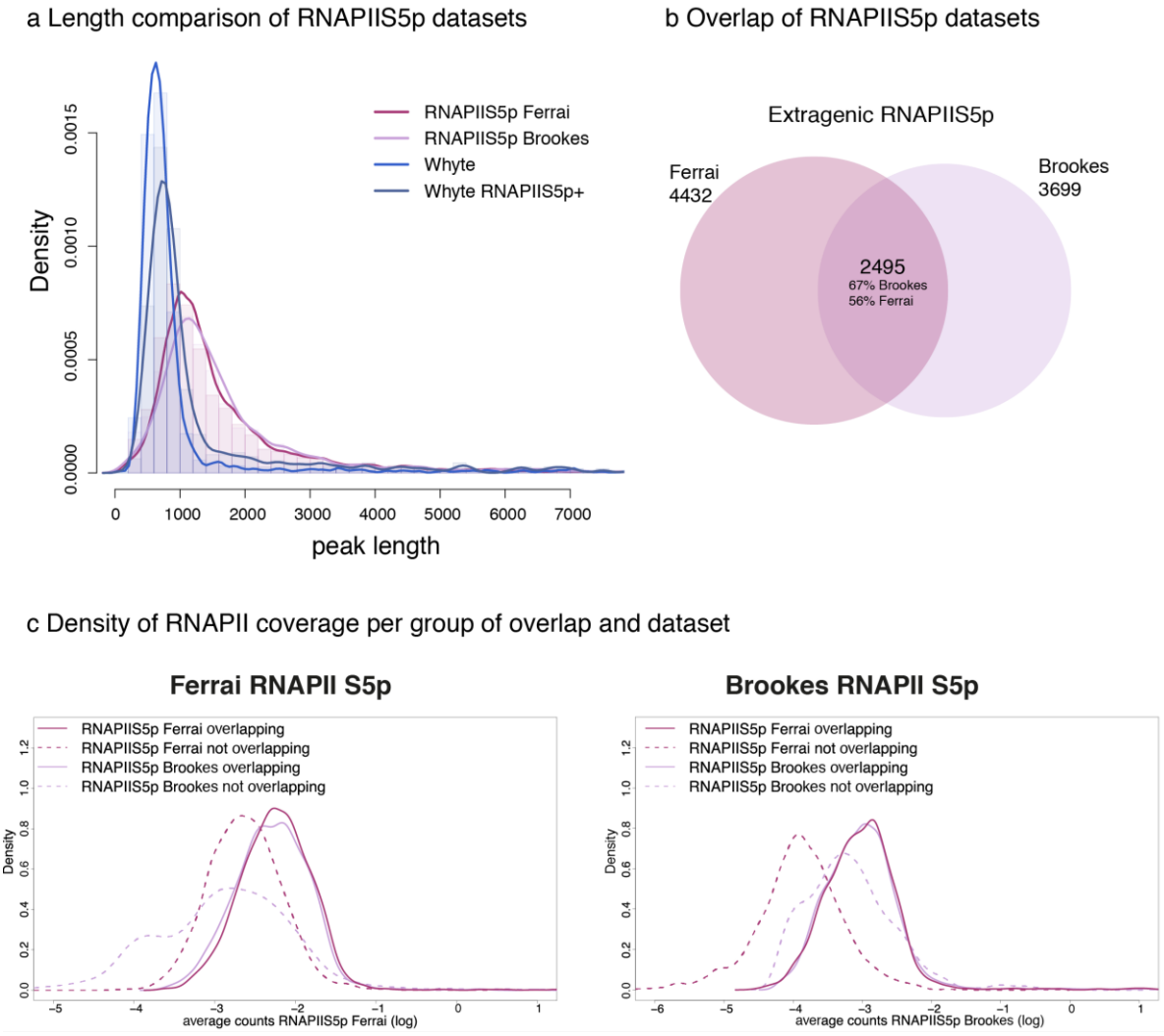


Fig 5.8: Comparison between Ser5p datasets from Brookes and Ferrai (Day0) datasets. a) Comparison of length distribution between RNAPIIS5p extragenic peaks from Brookes dataset, Ferrai day0 datasets, extragenic Whyte enhancers, and extragenic Whyte enhancers with RNAPIIS5p. b) Venn diagram of the overlap of Brooked and Day0 RNAPIIS5p datasets at extragenic regions. c) On the left: density distribution of RNAPII5p day0 reads at extragenic RNAPII5p regions from Day0 (dark pink) and Brookes (light pink). Solid lines represent regions overlapping between datasets, dotted lines represent regions not overlapping within datasets. On the right: density distribution of RNAPII5p Brookes reads at extragenic RNAPII5p regions from Day0 (dark pink) and Brookes (light pink). Solid lines represent regions overlapping between datasets, dotted lines represent regions not overlapping within datasets.

To understand whether regions shared by the two datasets were more strongly associated with RNAPIIS5p, I examined its enrichment in the common and unique regions. Regions overlapping between the two datasets are the most enriched for RNAPIIS5p (Fig 5.8c). In contrast, regions not overlapping show a lesser degree of enrichment. It is interesting to notice that regions specific for the Ferrai dataset (46C) have a higher enrichment for RNAPIS5p from Ferrai (46C)

compared to the ones of the Brookes dataset (OS25), and vice versa. This suggests that the differences in extragenic RNAPII highlight biological differences and are not easily explained by different sequencing depths of the two datasets or by technical noise in the detection of RNAPIIS5p extragenic regions. Similar analysis for RNAPIIS7p, RNAPIIS2p, and H3K27me3 enrichment can be found in Appendix Fig 5.A1. It is possible that highly enriched and shared regions are more stably bound by RNAPII, while the specific regions of each dataset can be more related to biological differences between the mESC clones in their signalling and metabolic pathways.

5.4.7 Comparison between extragenic RNAPIIS5p regions and published enhancer regions

To understand whether extragenic RNAPIIS5p regions mark putative enhancers, I compared the extragenic RNAPIIS5p regions with published enhancer datasets and analysed their chromatin marks. Only ~30% of extagenic RNAPIIS5p regions co-localise with either Whyte or Cruz Molina enhancers (Fig 5.9a), which is however in line with the low co-localisation of different published enhancer lists shown in chapter 3 of the current thesis and in previous publication (Benton et al., 2017). However, most (94%) of the regions that do not overlap Whyte or Cruz Molina enhancer lists, overlap with known enhancer marks; H3K27ac, P300, H3K4me1 and the non-canonical H3K122ac (Fig 5.9b). Regions co-localising with published enhancers show the same pattern, with a preference for co-localising with all the enhancer marks considered, whereas the same was true for only the minority of extragenic RNAPIIS5p regions not overlapping with one of the two enhancer lists. Similar comparisons at enhancer regions precedently identified not overlapping with RNAPII reveal a similar pattern of co-association with all considered marks (Appendix Fig 5.A2).

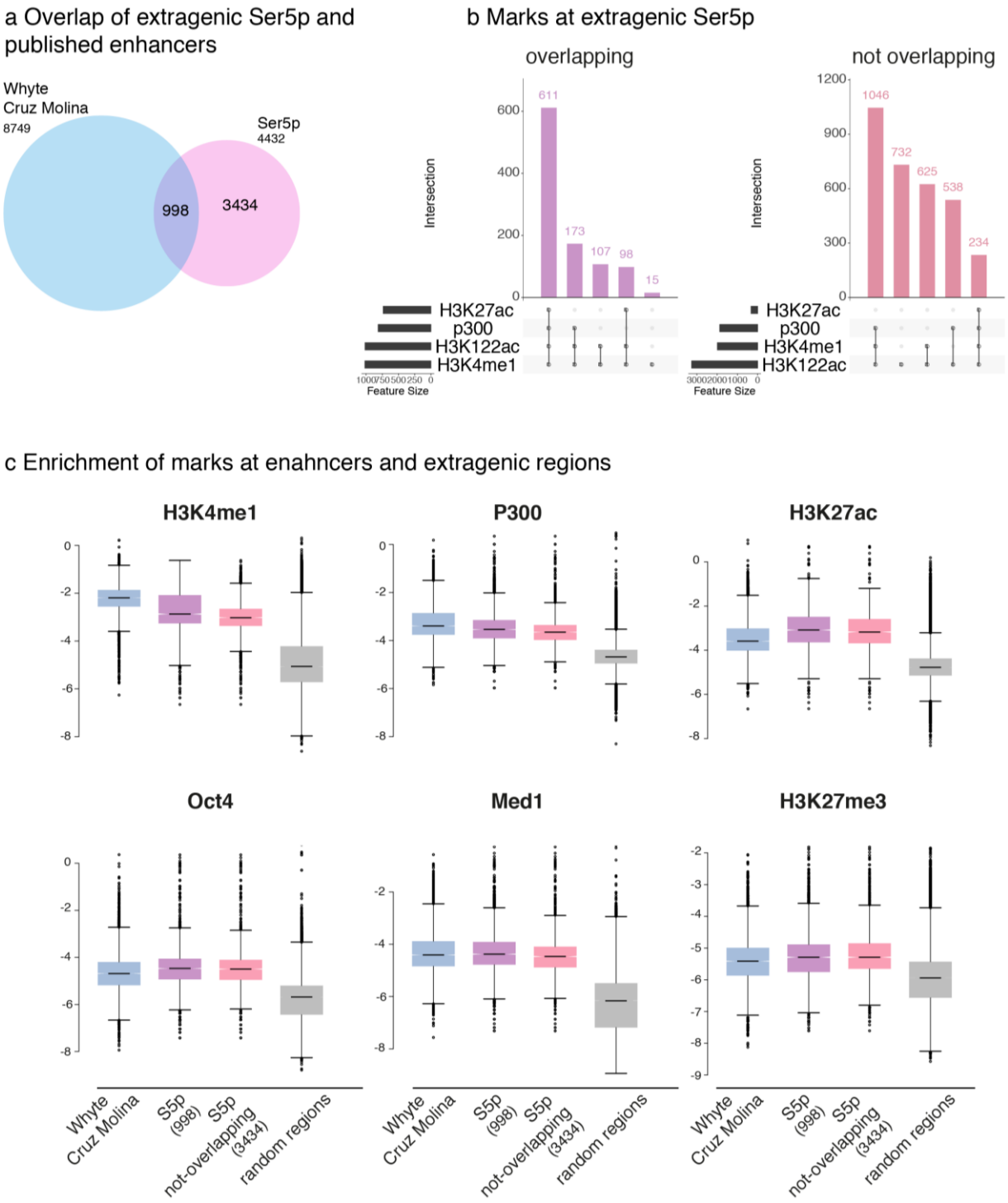


Fig 5.9: Extragenic RNAPII regions identify regions with enhancer marks. a) Overlap between extragenic RNAPIIS5p regions and Whyte and Cruz Molina extragenic enhancers. b) Combinations of enhancer marks at extragenic RNAPII5p regions co-localising with Whyte and Cruz Molina extragenic enhancers (on the left, in light purple) and not co-localising with Whyte and Cruz Molina extragenic enhancers (on the right, in pink). c) Enrichment of selected enhancer marks at Whyte and Cruz Molina enhancers (in blue), extragenic RNAPIIS5p regions co-localising with Whyte and Cruz Molina extragenic enhancers (in light purple), and extragenic RNAPIIS5p regions not co-localising Whyte and Cruz Molina extragenic enhancers (in pink). Extragenic random regions are shown as reference (in grey).

To better dissect differences and similarities between the two types of RNAPII extragenic region and understand their chromatin state, I calculated their enrichment for different enhancer markers (H3K4me1, P300, H3K27ac, Oct4, Med1 and H3K27me3). In general, RNAPIIS5p extragenic

regions have a comparable enrichment for enhancer features to published enhancers, which are above enrichment at random regions, without noticeable differences whether they co-localise (labelled S5p) or not with published enhancer lists (Fig 5.9c). H3K4me1 is slightly reduced at RNAPIIS5p extragenic regions, especially the regions not co-localising with published enhancers. The reason could lay in the fact that these regions could be transcribing and therefore enriched in H3K4me3 over H3K4me1. Interestingly, RNAPIIS5p extragenic regions tend to have higher enrichment for H3K27ac, a mark of active enhancers, higher Oct4 and H3K27me3 enrichment, compared Whyte and Cruz Molina enhancer regions.

Taken together these results demonstrate that extragenic RNAPIIS5p regions are enriched for known active and poised enhancer marks, and the enrichment of these marks is not dependent on whether these regions are identified by other approaches.

5.4.8 Classification of RNAPII states at extragenic regions

To study the activation state of extragenic RNAPIIS5p regions, I classified them according to their co-occupancy in RNAPIIS7p, RNAPIIS2p, and H3K27me3. The classification strategy applied was the same performed for extragenic Whyte enhancers (in Chapter 4, section 4.4.3). Briefly, the enrichment for each mark was calculated in the extragenic RNAPIIS5p regions. A 5% false positive cut-off was calculated over the random enrichments and used to classify as positive or negative for the feature of interest at the extragenic RNAPIIS5p regions. Positive regions were enriched above the cut-off and overlap with a BCP peak of the classifying feature (Fig 5.10a).

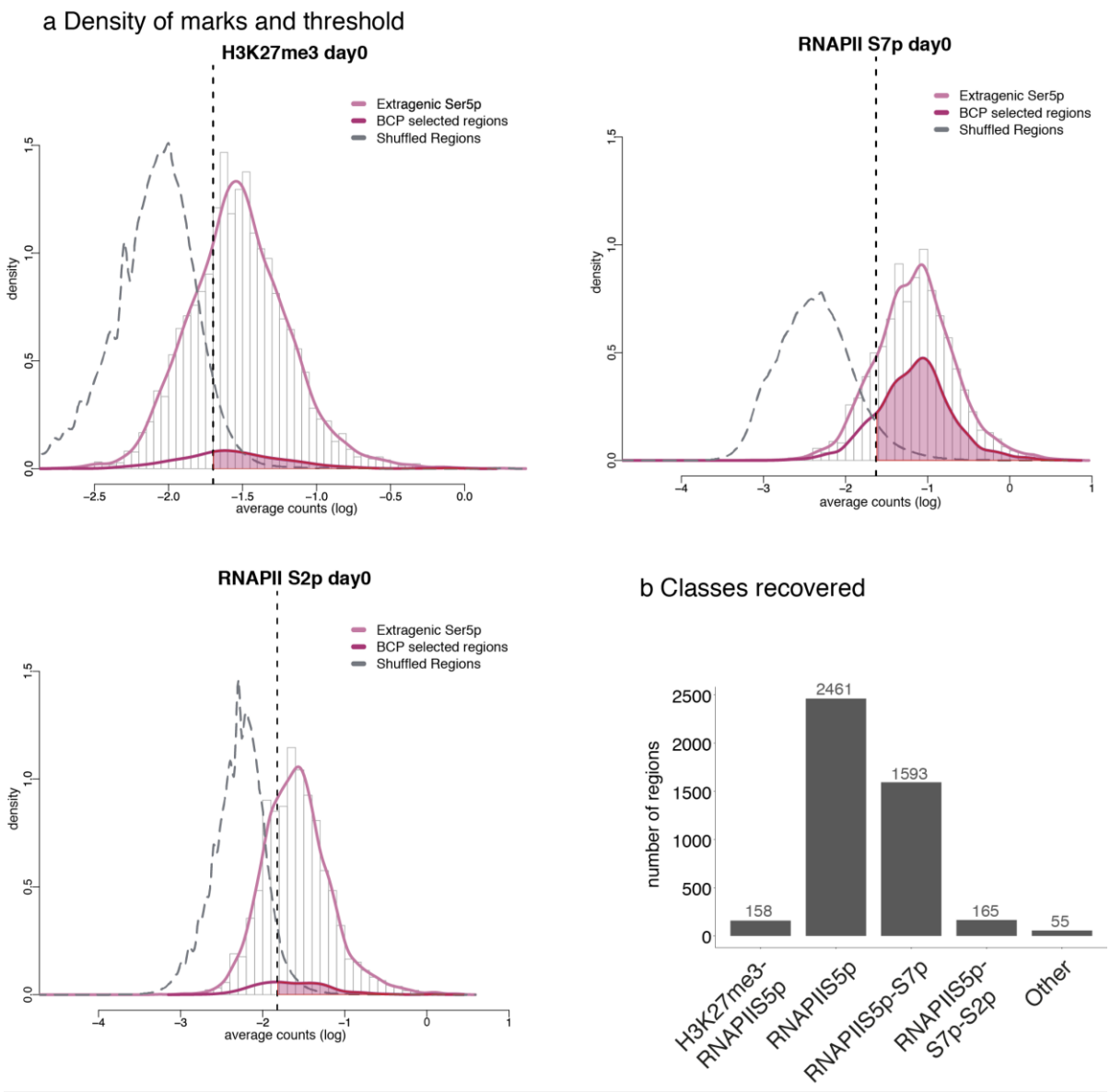


Fig 5.10: Classification of RNAPII states extragenic RNAPIIS5p regions from Day0 dataset. a) Classification of H3K27me3, RNAPIIS7p and RNAPIIS2p at RNAPIIS5p extragenic regions. Thin purple lines: enrichment of the considered feature at all extragenic RNAPIIS5p regions; thick purple lines: enrichment of the considered feature at extragenic RNAPIIS5p region overlapping a peak of the featured considered; grey dotted curved line: enrichment at extragenic random regions. Vertical dotted line: 5% FP cut-off. b) Number of RNAPII classes recovered.

To investigate the coincidence of the four markers considered, I overlapped their coordinates, and found that the most represented configuration of the extragenic RNAPIIS5p regions is RNAPIIS5p alone (2461), followed by RNAPIIS5p-S7p (1593), RNAPIIS5p-S7p-S2p (165) and H3K27me3-RNAPIIS5p (158) (Fig 5.10b). A small number (35) of regions had other minor combinations of markers and were not considered in the following analyses.

In conclusion, RNAPII is found in different activations states at extragenic regions, from poised together with Polycomb to fully elongating, similarly to the results on published enhancer lists presented in chapter 4.

5.4.9 RNAPII is more enriched at more active extragenic regions

To investigate the distribution and level of occupancy of RNAPII and Polycomb mark H3K27me3 at RNAPIIS5p extragenic regions, I analysed their pattern of enrichment. RNAPII modifications are enriched in the peak area concordantly with their classification (Fig 5.11a). H3K27me3 spreads over the RNAPII extragenic region and its surroundings, potentially generating a wide repressive area, as is known for its occupancy at Polycomb repressed genes (Bernstein et al., 2006a). Interestingly, RNAPIIS5p is enriched at the same level at regions with or without Polycomb. In contrast with H3K27me3 and RNAPIIS5p, RNAPIIS7p and RNAPIIS2p show a major peak of enrichment in the center of the extragenic RNAPIIS5p region. Importantly, no detectable enrichment is found at the extragenic RNAPIIS5p regions considered in a control ChIP experiment performed in the absence of primary antibody (Mock IP control).

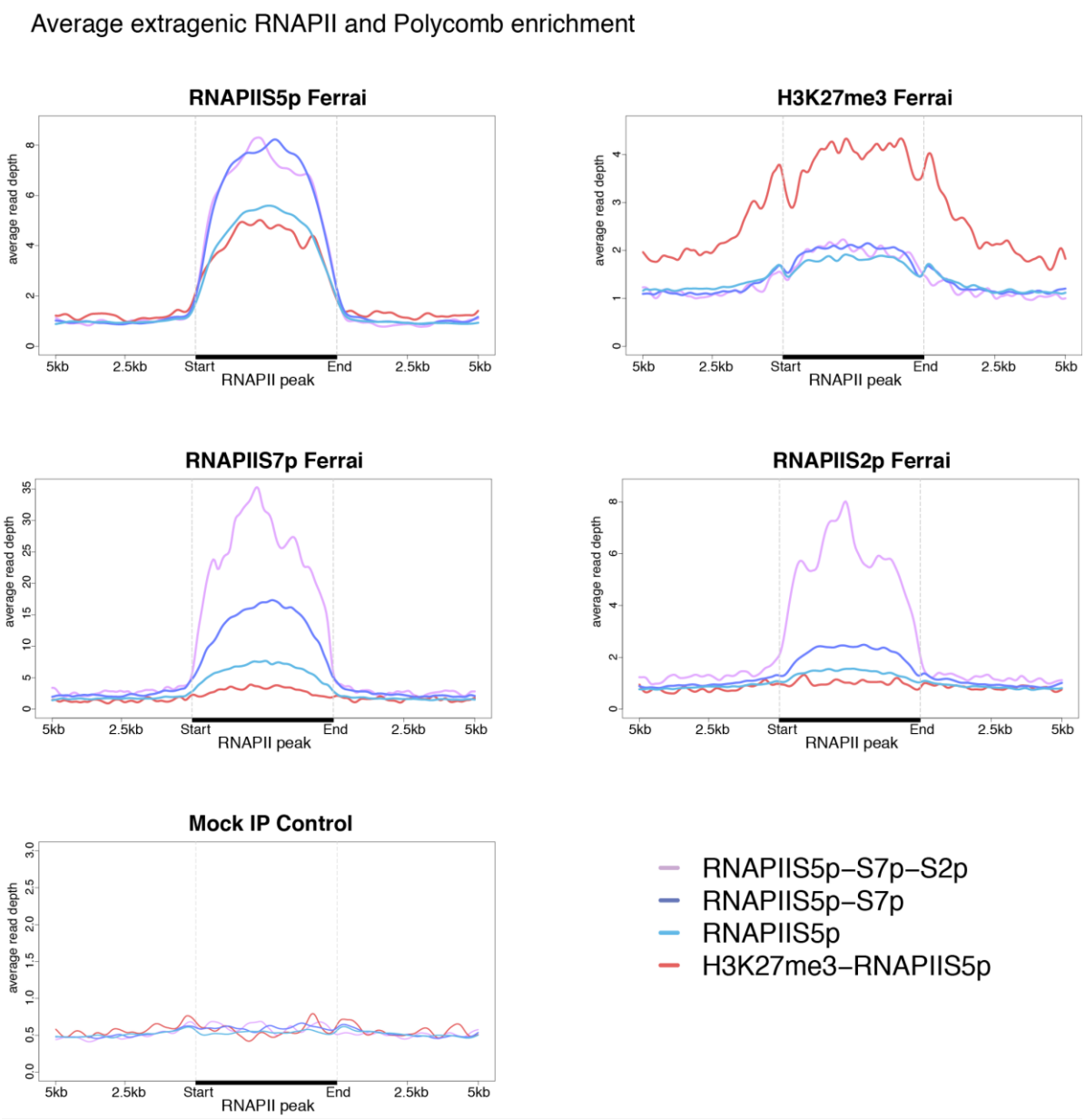


Fig 5.11: RNAPII modifications are enriched at extragenic RNAPII regions concordantly with their classification. a) Average enrichment of feature used in the classification per classes of extragenic RNAPII regions. Colorcoded are the classes of extragenic RNAPII regions. The average enrichment inside the regions +/- 5kb is shown. Regions shorter than 1kb were not included in the averaging.

5.4.10 Extragenic RNAPII regions marked by Polycomb are closer to repressed genes

Next, I explored the proximity of each class of extragenic RNAPIIS5p region to genes, and the promoter state of these genes. For example, H3K27me3-RNAPIIS5p regions could be located closer to Polycomb repressed genes, than extragenic regions containing S7p and S2p, to build a repressive chromatin environment, as was previously suggested for poised enhancers (Koenecke et al., 2017; Rada-Iglesias et al., 2011). Most extragenic RNAPII regions were found located at more than 10kb from the closest annotated gene >75% of the times, with most at more than 50kb

(>30% per class; Fig 5.12a). Interestingly, H3K27me3-RNAPIIS5p regions are the closest to genes with distances of <10kb ~20% of the time.

Fig 5.12: Extragenic H3K27me3 positive regions are closer to Polycomb repressed genes. : a) Gene distance of extragenic RNAPIIS5p regions from the closest gene, divided in bins. b) Scheme of states of promoters considered in this analysis. Promoters states were define in Ferrai et al. 2017 .c) State of the closest gene per class of extragenic RNAPII. d) Distance per class of closest gene, divided per class of enhancer and gene states.

To understand whether H3K27me3-RNAPIIS5p reside closer to Polycomb repressed genes, I took advantage of a published list of promoter states derived from the same datasets (Ferrai et al., 2017). In this list, active promoters were defined as positive for RNAPIIS5p, RNAPIIS7p and expressing mRNA; PRC-RNAPIIS5p promoters were defined as positive for H3K27me3

and RNAPIIS5p and negative for the other features considered; PRC promoters were defined as positive for H3K27me3 and negative for the other features considered; Inactive promoters were defined as negative for all the features considered (Fig 5.12b). I decided to consider both PRC and PRC-RNAPIIS5p regions to investigate whether a difference could be found in respect to extragenic RNAPII position, which could be indicative of different regulatory mechanisms.

Interestingly, I find that the state of the closest gene to H3K27me3-RNAPIIS5p regions is often Polycomb repressed (38% of closest genes) (Fig 5.12c), with a preference for H3K27me3-RNAPIIS5p compared to other extragenic RNAPII regions. Moreover, H3K27me3-RNAPIIS5p regions are found between 2-50kb closer to Polycomb-repressed genes than to active genes (2-50kb 28% Polycomb-repressed, 14% active; 2-10kb 11%, 2%). This result raises the possibility that H3K27me3-RNAPIIS5p extragenic regions reside close to repressed genes to help keep or promote a repressive environment. This analysis also shows that a clear picture of enhancer and promoter state correlation is challenging when only linear distance is taken into account.

5.4.11 Extragenic RNAPII regions co-associate with active histone modifications and transcription factors

To investigate whether different state of RNAPII at extragenic regions associate with specific enrichments of histone modifications and TFs linked with enhancer activity, I analysed the enrichment of selected modifications and TFs in each region according to the class of extragenic RNAPIIS5p regions (Fig 5.13). H3K4me1 spreads in the surroundings of extragenic RNAPII regions of all classes, while marks related with activity, such as H3K27ac and P300 are more enriched the more active is the state of RNAPII, and are more centered in their distribution. Interestingly, TFs such as Nanog, Sox2, and Oct4 are enriched to different degrees inside the extragenic RNAPII classes. This could reflect the partial co-localisation between extragenic RNAPII regions and Whyte enhancers (Fig 5.9a), and show that some regions occupied by RNAPII are not bound by these specific TFs. Nevertheless, it is important to notice that Brd4 and Med1 are present at extragenic RNAPII regions. Interestingly, proteins involved in chromatin looping, such as Caph2 and Nipbl are also present at all the classes, while CTCF is depleted in all regions. (For a quantification of the enrichment levels of H3K27me3, H3K27ac, H3K4me1, H3K122ac, Klf4, Nanog, Sox2, Oct4, Smad1, Caph2, CTCF, Nipbl, Med1 please refer to Appendix Fig 5.A3).

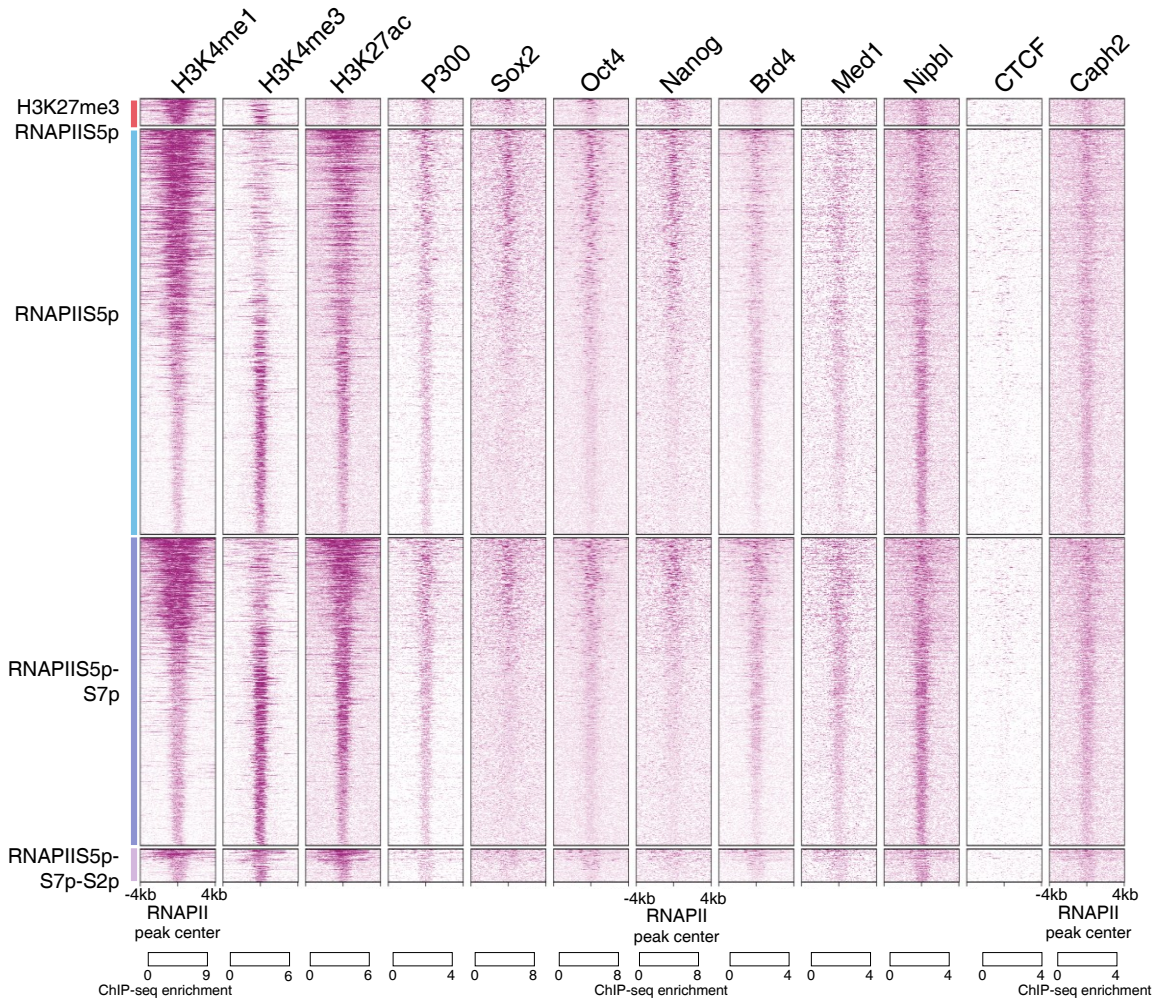


Fig 5.13: Extragenic RNAPII classes show diverse enrichment for chromatin marks. Positional heatmaps showing the positional enrichment of selected marks at RNAPII regions. Regions +/- 4kb from the middle point of the extragenic RNAPII region are shown. On the bottom scale per feature are specified.

Altogether, the results show how extragenic RNAPII can capture different chromatin states. Extragenic RNAPII regions are enriched in enhancer marks and also in chromatin looping associated proteins, which would suggest a role in 3D gene regulation.

5.4.12 Datasets used to study transcriptional activity at extragenic RNAPIIS5p regions

To study the regulation of transcription at extragenic RNAPII regions, I explored both RNA-seq and ChIP-seq datasets, published and unpublished (a list of the datasets used in the following paragraphs can be found in Tables 5.2 and 5.3, for RNA-seq and ChIP-seq datasets, respectively).

Table 5.2: RNA-seq datasets used in the current chapter. RNA-seq datasets used in the current chapter. *Processed by Elena Torlai Triglia. § Processed by Robert A Beagrie.

Datasets	mESC clone	Source	GEO
polyA RNA*	46C	Ferrai *et al.*, 2017	GSM2474132
polyA RNA	OS25	Morris, JK *et al.*, unpublished	-
Nascent RNA §	46C	Morris, JK *et al.*, unpublished	-
Nascent RNA	OS25	Morris, JK *et al.*, unpublished	-
Total RNA	46C	Fernandes, AM *et al.*, unpublished	-
Exosome KD si (totalRNA)	OS-25	Morris, JK *et al.*, unpublished	-
Exosome KD ctrl (totalRNA)	OS25	Morris, JK *et al.*, unpublished	-
Flavopiridol DMSO (nascent RNA)	OS25	Morris, JK *et al.*, unpublished	-
Flavopiridol KD (nascent RNA)	OS25	Morris, JK *et al.*, unpublished	-

Table 5.3: ChIP-seq datasets used for the transcriptional analysis. ChIP-seq datasets of regulators of transcription or perturbation of transcription used in the current chapter.

Datasets	mESC_clone	Type	Source	GEO
Cdk8	V6.5	Transcription regulator	Reddy *et al.*, 2014	GSM1463943
Cdk7	V6.5	Transcription regulator	Reddy *et al.*, 2014	GSM1463942
Cdk9	V6.5	Transcription regulator	Whyte *et al.*, 2013	GSM1082347
Spt5	V6.5	Transcription regulator	Rahl *et al.*, 2010	GSM515370
Dis3*	46C	Exosome subunit	Fernandes, AM *et al.*, unpublished	-
Erk2*	E14	Transcription regulator	Tee *et al.*, 2014	SRR953607
Nelfa	V6	Transcription regulator	Rahl *et al.*, 2010	GSM515366
N20_1 wt	E14	Total RNAPII	Tee *et al.*, 2014	SRR953613
N20_2 wt	E14	Total RNAPII	Tee *et al.*, 2014	SRR953614
N20 Erk KO	E14	Total RNAPII	Tee *et al.*, 2014	SRR953602
Ser5p wt	E14	RNAPII modification	Tee *et al.*, 2014	SRR953615
Ser5p Erk KO	E14	RNAPII modification	Tee *et al.*, 2014	SRR953603
Input wt	E14	-	Tee *et al.*, 2014	SRR953585

RNA-seq datasets cover different maturation states of RNA (Total RNA-seq, Nascent RNA-seq, PolyA RNA-seq), perturbation of transcription (Nascent RNA-seq after Cdk9 inhibition with Flavopiridol), and RNA maturation surveillance (Total RNA-seq after Exosome KD). ChIP-seq datasets include mapping of kinases involved in RNAPII regulation (Cdk9, Erk2), pausing

factors (Nelfa, Spt5), Exosome (Exosome catalytic subunit Dis3), co-factors involved in Polycomb repression (Utf1) and RNAPII modification in WT and after Erk2 inhibition (total RNAPII and RNAPIIS5p ChIP-seq).

5.4.13 RNAPII at extragenic regions transcribe differently mature RNAs

Different states of RNAPII are associated with different stages of the transcription cycle. RNAPII marked by S5p, S7p and S2p is associated with productive elongation that leads to fully mature mRNA (Brookes and Pombo, 2009), while poised RNAPIIS5p, in the absence of S7p and S2p, leads to abortive transcription (Jia et al., 2012). eRNAs, on the other hand, were shown to be unstable and not polyadenylated (Andersson et al., 2014a; Arner et al., 2015; Flynn et al., 2011), although they are capped, but it is not clear which phosphorylation state of RNAPII is most associated with eRNA production. Moreover, promoters can have divergent transcription which produces PROMoter uPstream Transcripts (PROMPTs), which are differently regulated from RNA originating from genes (Almada et al., 2013; Flynn et al., 2011), and resemble eRNAs. For example, PROMPTs and eRNAs were shown to be under exosome surveillance (Andersson et al., 2014b; Flynn et al., 2011).

To investigate whether extragenic RNAPII regions produce differently mature RNAs based on their RNAPII state, I analysed different RNA-seq datasets. Total RNA-seq captures all the RNA species in a cell, except for the rRNAs that were depleted prior to sequencing. Nascent RNA-seq (RNAChIP) captures RNA associated with RNAPIIS5p bound to chromatin, and are enriched for nascent RNAs (Kelly J. Morris, personal communication). PolyA RNA-seq is the pool of mature RNAs, with a polyA tail. I compared the amount of RNA produced at the extragenic RNAPIIS5p regions with the expression at active genes (on exons) and also with the signal at upstream regions of active genes (PROMPTs).

Differently classified RNAPII extragenic regions transcribe different amounts of total RNA, from mostly not detectable at H3K27me3-RNAPIIS5p to levels higher than at PROMPTs at RNAPIIS5p-S7p-S2p regions. Specifically, total RNA is detected at regions with RNAPIIS5p alone to almost similar levels to PROMPTs, however when RNAPIIS2p is present the transcription is higher than 1 TPM in 62% of the regions (Fig. 5.14a). Strikingly, levels of Nascent RNA at RNAPIIS5p-S7p-S2p are comparable to active genes (Fig 5.14b), however this is not the case for polyA-RNA (Fig 5.14c), where the levels are significantly lower at extragenic regions compared to active coding regions. Similar data were obtained analysing RNA-seq

datasets originating from OS25 cells (Appendix Fig 5.A4, Methods Table 2.1 for details on the OS25 datasets).

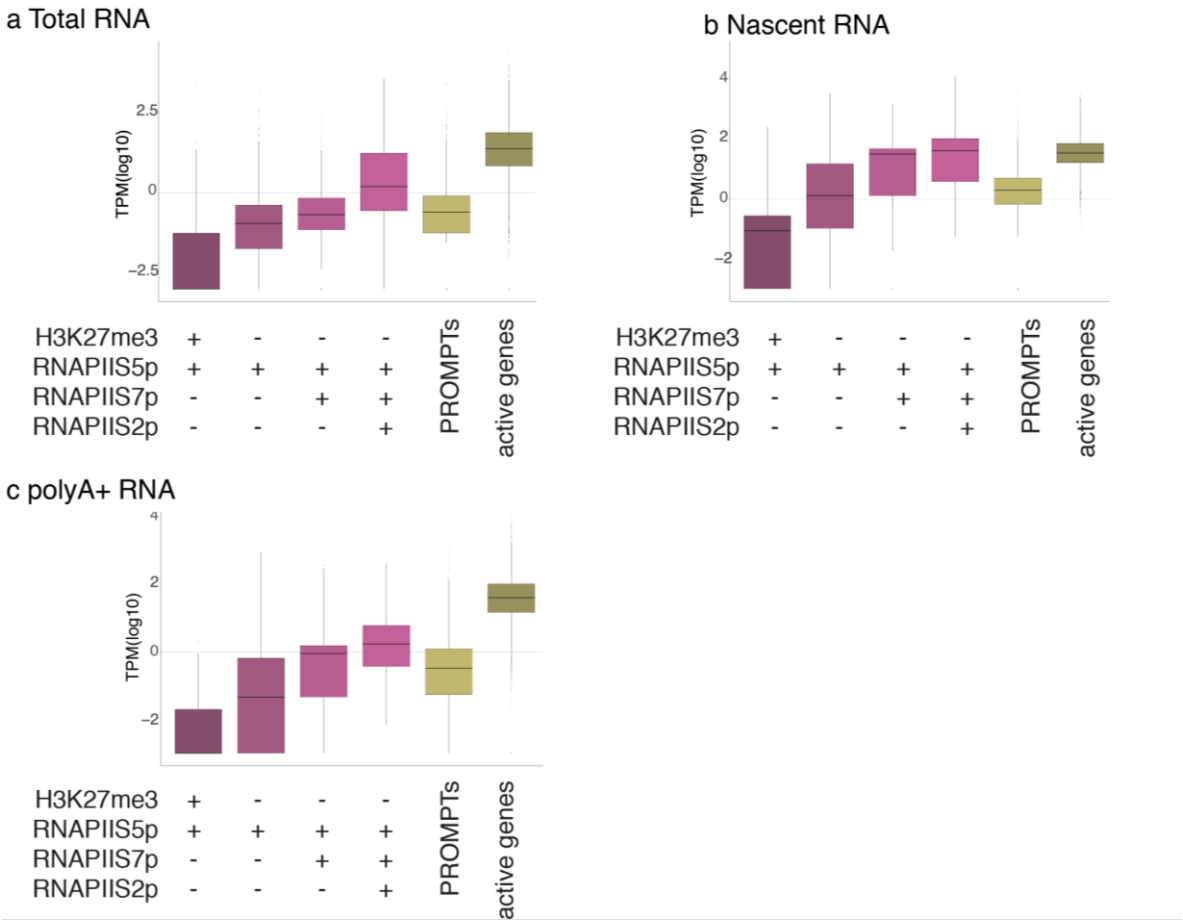


Fig 5 .14: Extragenic RNAPII regions transcribe differently mature RNAs depending on their activation state. a) total RNA. b) Nascent RNA (RNAChIP). c) polyA-RNA. Log10 TPM + pseudo count are shown for clarity. Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference.

These results show that RNAPII at extragenic regions transcribes RNAs with different degrees of maturation and transcription levels, which reflect RNAPII activation states. Interestingly, nascent RNA levels at active extragenic region are comparable to RNA levels at active genes and significantly higher than PROMPTs, while fully mature RNA is lowly expressed. This confirms that bonafide detection of RNAPII at extragenic regions and shows that transcription occurs at extragenic regions, leading to RNA molecules that are not processed to full maturation and/or degraded.

5.4.14 Extragenic RNAPII transcripts are under exosome surveillance

To test whether transcripts are actively degraded at extragenic RNAPII regions, I analysed Exosome occupancy at these regions. Exosome was described to be present at enhancer regions and degrade eRNAs (Andersson et al., 2014b; Flynn et al., 2011). Interestingly, H3K27me3-

RNAPIIS5p regions are the most enriched for Dis3, a catalytic subunit of the Exosome machinery (Fig 5.15a), with comparable levels of occupancy at Polycomb promoter regions. Exosome is known to be active to prevent PROMPTs accumulation at transcribed upstream regions of active genes (Flynn et al., 2011), but it is unclear whether Exosome activity at active genes coincides with chromatin association that could be detected by ChIP.

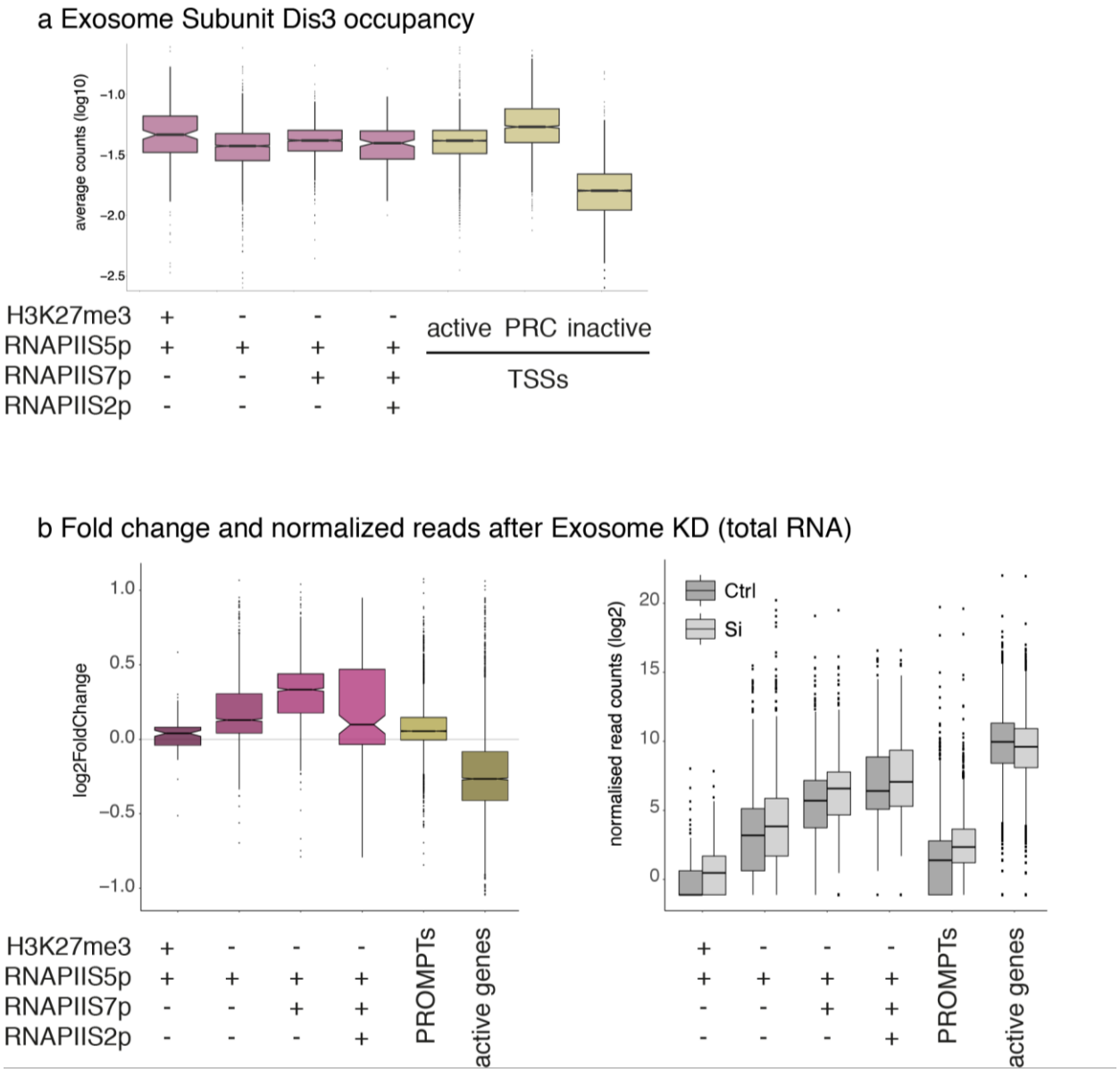


Fig 5.15: Extragenic RNAPII regions are under exosome surveillance. a) Average count enrichment of exosome catalytic subunit Dis3 at extragenic RNAPII regions and at TSS of differently active promoters. Log10 of normalised reads per length + pseudocount are shown for clarity. b) Foldchange (on the left) and normalised reads (on the right) of total RNA levels after Exosome KD. Ctrl: cells treated with a control Si. Si: cell treated with Si against exosome. Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference.

To understand whether Exosome is actively involved in the degradation of transcripts produced from extragenic RNAPII regions, I took advantage of a total RNA-seq dataset previously produced in the lab (KJ Morris *et al.*, in preparation). The total RNA-seq was performed in

control and after Exosome knockdown using siRNAs. The depletion of Exosome leads to an increase in the levels of total RNA detected at extragenic RNAPII regions and a low effect on PROMPTs (Fig 5.15b).

In conclusion, RNAs produced at extragenic RNAPII regions are under Exosome surveillance.

5.4.15 Regulators of transcription at coding regions are also enriched at extragenic RNAPII regions

To investigate whether the transcriptional activity of extragenic RNAPII is regulated similarly to what happens at genes, I analysed the occupancy of known transcriptional regulators at extragenic RNAPIIs5p regions (Fig. 5.16a,b), including: Spt5 and Nelfa, involved in the promoter proximal pausing; Cdk9 and Erk2, two kinases that phosphorylate CTD at active and Polycomb genes, respectively; and Utf1, a co-factor that regulates Polycomb repressed genes.

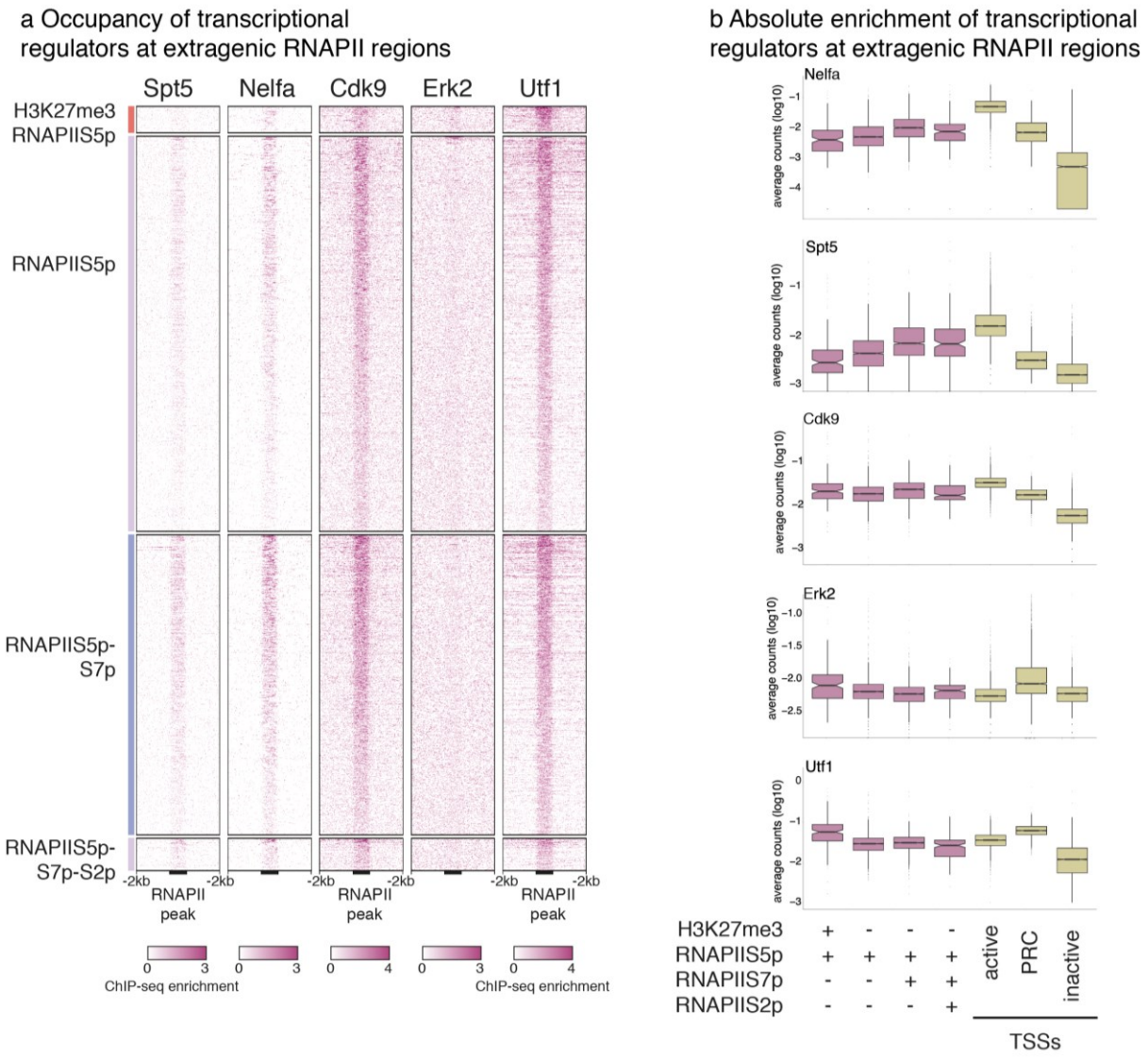


Fig 5.16: Transcription regulators are differentially enriched at extragenic RNAPII classes. a) Positional enrichment of selected transcriptional regulators at extragenic RNAPII regions. Extragenic RNAPII regions +/- 2kb are shown. On the bottom scale per feature are specified. b) Absolute enrichment of selected features at extragenic RNAPII region and differently active TSSs.

Spt5 and Nelfa are more enriched at extragenic RNAPIIS5p regions with increased activation of RNAPII, and these two factors do not spread in the surrounding chromatin, but are confined to the RNAPII peak (Fig 5.16a). Spt5 has similar enrichment at RNAPIIS5p-S7p and RNAPIIS5p-S7p-S2p extragenic regions, while Nelfa is more enriched at RNAPIIS5p-S7p than at RNAPIIS5p-S7p-S2p regions. These observations are in accordance with the promoter proximal pause release mechanism described at genes, where Nelfa and Spt5 bind and cause RNAPII pausing. After phosphorylation by Cdk9, Nelfa dissociates releasing paused RNAPII, while Spt5 remains associated to elongating RNAPII throughout the gene body (Adelman and Lis, 2012). As expected from Erk2 roles at Polycomb-repressed promoters (Tee et al., 2014), Erk2 shows its highest enrichment at H3K27me3 positive regions, although barely detectable. In contrast, Cdk9 and Utf1 are present at all RNAPIIS5p extragenic regions considered. A comparison between the

enrichment levels at extragenic regions and TSSs shows lower occupancy of Nelfa, Spt5 and Cdk9 at extragenic regions compared to coding regions (Fig 5.16b).

In conclusion, extragenic RNAPII regions are bound by known transcriptional regulators. Erk2, a regulator of RNAPII posing which are specific for Polycomb genes is more enriched at H3K27me3 positive regions. Taken together, these results suggest that transcription at extragenic regions is regulated similarly to genes.

5.4.16 RNAPII transcription is sensitive to Cdk9 inhibition

To investigate Cdk9 presence at extragenic RNAPIIS5p regions in more detail, I analysed Nascent RNA dataset from control cells and matched cells treated with Flavopiridol (1h; 10 uM), which inhibits Cdk9. Cdk9 phosphorylates the RNAPII-CTD, Nelfa pausing factor and Spt5 positive elongation factor and regulates the transition between initiation to elongation stages of RNAPII transcription at coding genes. Inhibition of Cdk9 with Flavopiridol blocks the transition between initiation to elongation and results in transcription inhibition.

Flavopiridol treatment results in decrease in nascent RNA levels at all extragenic RNAPII regions considered, with a stronger effect for RNAPIIS5p-S7p and RNAPIIS5p-S7p-S2p (Fig 5.17a), which are the more transcribed extragenic RNAPII regions. A somehow unexpected result is the increase in nascent RNA from the coding regions of active genes which is probably due to RNAPII stalling at TESs for inefficient termination (Jonkers et al., 2014). These results show that RNAPII transcription at extragenic regions is under Cdk9 control.

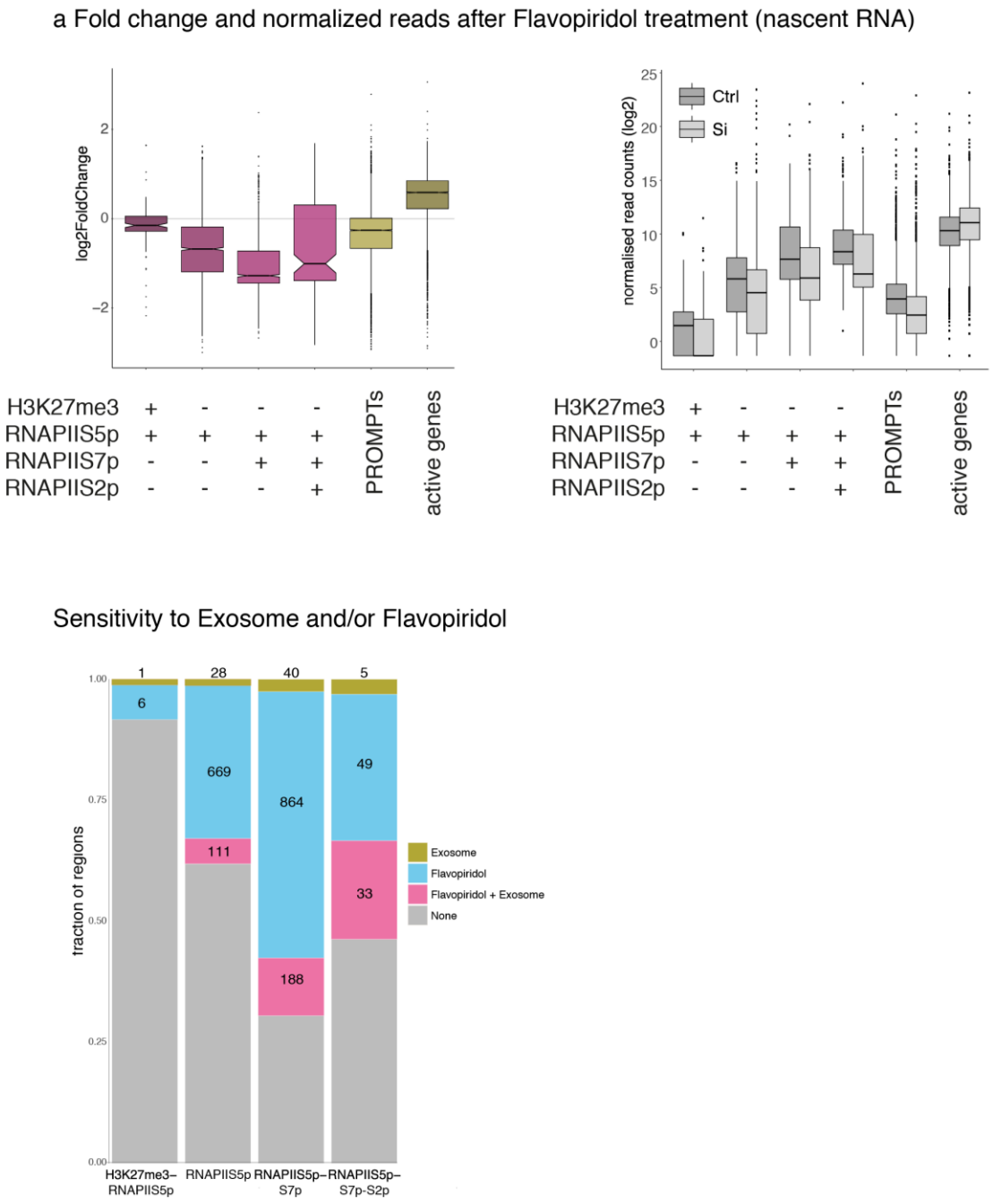


Fig 5.17: Extragenic RNAPII regions are sensitive to Cdk9 inhibition. a) Foldchange (on the left) and normalised reads (on the right) of nascent RNA levels after Flavopiridol treatment. Ctrl: untreated cells. Si: cell treated with Flavopiridol (1h; 10 uM). Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference. b) Sensitivity of extragenic RNAPII regions to Flavopiridol treatment and Exosome KD. Dotplot showing all the extragenic RNAPII regions and their sensitivity to Flavopiridol (x-axis) and Exosome KD (y-axis). Log2 of normalised reads is shown. c) Sensitivity to Flavopiridol, Exosome KD or both per class. Fraction of regions per class is shown.

To understand whether regions under Cdk9 control are also under Exosome surveillance, I compared the fold change in the two different treatments at all the extragenic regions. Interestingly, regions bound by more active RNAPII are under Exosome and/or Cdk9 surveillance (Fig 5.17c, a plot showing the definition of sensitivity to Exosome or Flavopiridol can be found in Appendix Fig 5.A6). RNAPIIS5p-S7p is predominantly regulated by Cdk9

alone, while RNAPIIS5p-S7p-S2p is almost equally regulated through Cdk9 and Exosome together. Possibly, Exosome complex plays a more prominent role at more actively transcribed regions, where it is important to avoid accumulation of unnecessary RNAs.

This analysis hints at the possibility that transcription at enhancers is regulated through CTD modifications and pausing release, as well as through active degradation of the resulting non-coding transcripts. Similar results were obtain at the RNAPII-bound extragenic Whyte enhancers and partially presented in Appendix Fig 5.A6.

5.4.17 Erk2 knock out influences RNAPII binding at extragenic regions

Polycomb repressed genes are regulated by Erk2 which phosphorylates poised RNAPIIS5p complexes at developmental genes in ESCs (Tee et al., 2014). To understand whether Erk2 regulates extragenic transcription at poised H3K27me3-RNAPIIS5p extragenic regions, I analysed Erk2 enrichment at extragenic RNAPIIS5p regions and I also investigated RNAPII occupancy at extragenic RNAPIIS5p regions and at active, Polycomb-repressed, and inactive genes after Erk knock out (KO) in mESCs .

Erk2 is preferentially enriched at extragenic regions bound by Polycomb, and at coding genes (Fig 5.18a) and to a lesser extent at other extragenic RNAPII regions. While the levels of Erk and H3K27me3 vary between H3K27me3 positive and negative regions, as expected, the levels of RNAPIIS5p enrichment are comparable (Fig 5.18b), also as expected from the identification of similar levels of S5p at active and Polycomb repressed genes.

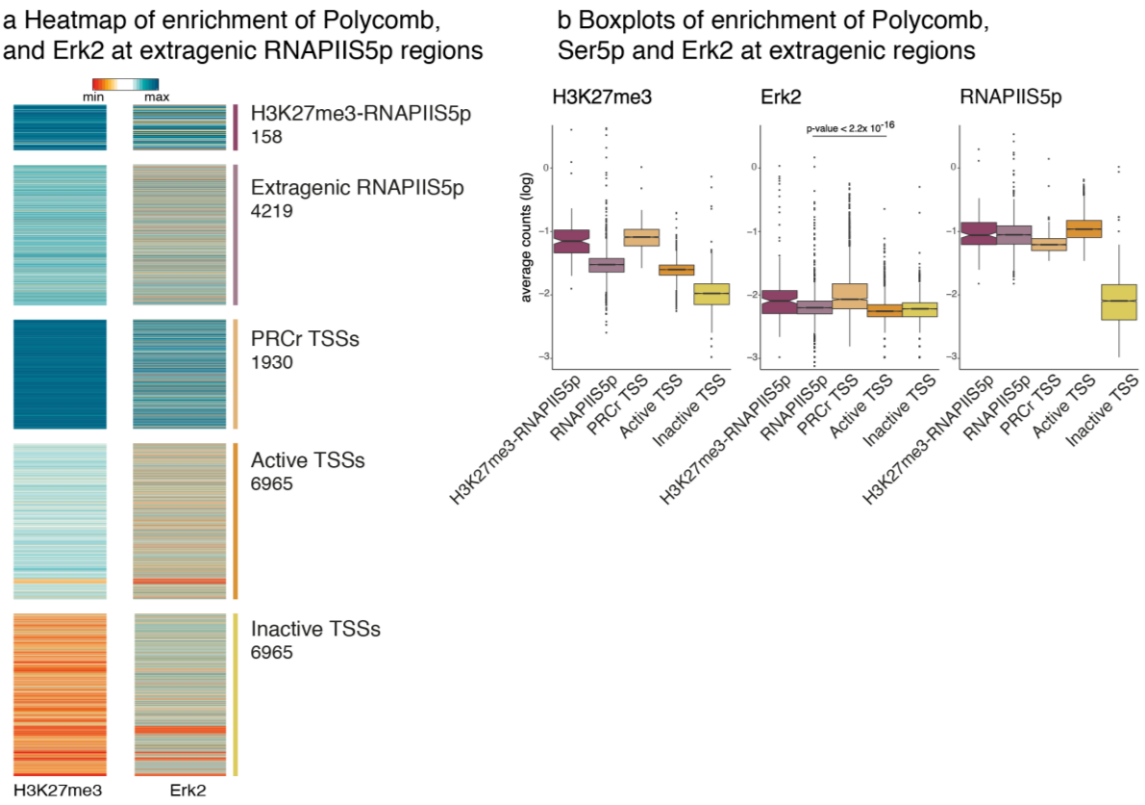


Fig 5.18: Erk2 is enriched at extragenic RNAPII regions. a) Heatmap showing the enrichment of H3K27me3 and Polycomb at extragenic RNAPII region with or without H3K27me3 and at Polycomb repressed genes, Active genes and Inactive genes, as reference. b) Absolute enrichment of H3K27me3, Erk2, RNAPIIS5p at extragenic RNAPII region with or without H3K27me3 and at Polycomb repressed genes, Active genes and Inactive genes, as reference. Log of average counts (normalised per length) + pseudo count are shown for better visualisation.

To test whether Erk2 regulates RNAPII occupancy at extragenic regions, I compared the enrichment of RNAPIIS5p and total RNAPII before and after Erk KO in mESCs. As previously observed at Polycomb-repressed genes marked by Erk2 (Tee et al., 2014), Erk KO decreases the occupancy of RNAPIIS5p and total RNAPII also at Polycomb-RNAPIIS5p extragenic regions (Fig 5.19a). As shown previously (Tee et al., 2014), active genes are not occupied by Erk2 and not influenced by Erk KO. In contrast, at extragenic regions RNAPIIS5p occupancy tends to decrease after Erk KO irrespectively of H3K27me3 (Fig 5.19b). Total RNAPII enrichment is also diminished at extragenic RNAPII regions without H3K27me3 (Fig 5.19c). The decrease at extragenic regions without Polycomb mark was an unexpected result that may inform about the mechanisms that recruit and regulate RNAPII at extragenic regions. To explore which RNAPII state at extragenic enhancers was also sensitive to Erk inhibition, I analysed all extragenic RNAPII classes and found that surprisingly the extragenic regions that loose RNAPIIS5p or total RNAPII occupancy are the ones positive for RNAPIIS2p (Fig 5.19d).

Figure 5.19

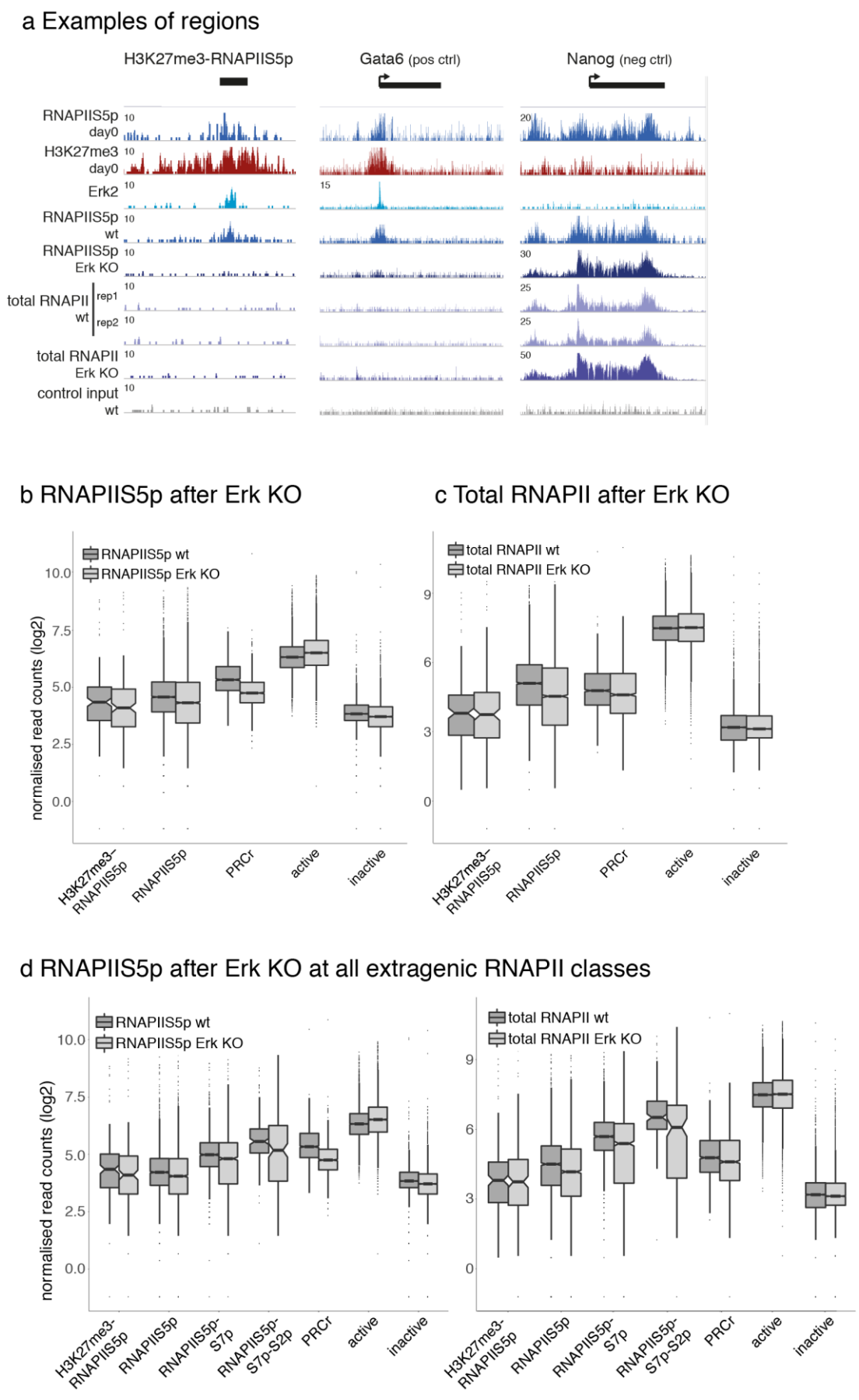

a Examples of regions
H3K27me3-RNAPIIS5p
Gata6 (pos ctrl)
Nanog (neg ctrl)
RNAPIIS5p day0
H3K27me3 day0
Erk2
RNAPIIS5p wt
RNAPIIS5p Erk KO
total RNAPII wt rep1
total RNAPII wt rep2
total RNAPII Erk KO
control input wt
b RNAPIIS5p after Erk KO
c Total RNAPII after Erk KO
RNAPIIS5p wt
RNAPIIS5p Erk KO
total RNAPII wt
total RNAPII Erk KO
normalised read counts (log2)
H3K27me3-RNAPIIS5p
RNAPIIS5p
PRCr
active
inactive
d RNAPIIS5p after Erk KO at all extragenic RNAPII classes
RNAPIIS5p-S7p
RNAPIIS5p-S7p-S2p

Fig 5.19: Erk KO reduces RNAPII binding at extragenic regions. a) Single region view of extragenic RNAPII regions (left), Polycomb repressed gene Gata6 (center), and active gene Nanog (right). Tracks of Erk2, H3K27me3, RNAPIIS5p before and after Erk KO and total RNAPII before and after Erk KO are shown. Control (Input) is also shown. Tracks always show 0 and a smoothing function of 2 was applied. Images generated with the IGV software. b) RNAPIIS5p absolute levels before and after Erk KO at extragenic RNAPII region with or without H3K27me3 and at Polycomb repressed genes, Active genes and Inactive genes. Log of normalised read counts calculated with DESeq2 are shown. c) total RNAPII absolute levels before and after Erk KO at extragenic RNAPII region with or without H3K27me3 and at Polycomb repressed genes, Active genes and Inactive genes. Log of normalised read counts calculated with DESeq2 are shown. d) RNAPIIS5p and total RNAPII absolute levels before and after Erk KO at extragenic RNAPII classes and at Polycomb repressed genes, Active genes and Inactive genes. Log of normalised read counts calculated with DESeq2 are shown.

In conclusion, Erk2 is bound preferentially at H3K27me3-RNAPIIS5p extragenic regions, but its knockout perturbs RNAPII binding at all extragenic RNAPII regions, especially very active ones, without affecting RNAPII occupancy at active genes.

These results suggest that RNAPII regulation at extragenic regions shares many of the features of regulation at coding regions.

5.5 Discussion

RNAPII marks putative enhancer regions, which are missed by other approaches. Different states of extragenic RNAPII have diverse enrichment for transcription factors, histone modifications, and proteins related with chromatin architecture. Moreover, extragenic RNAPII transcribes RNAs with different levels of maturation that are degraded by the Exosome machinery. Transcription at enhancers is also regulated by kinases that act on the CTD and on transcription pausing factors.

5.5.1 Extragenic RNAPII marks putative regulatory regions

RNAPII was previously described to bind enhancers in macrophages upon stimuli (De Santa et al., 2010) and poised enhancers when they get activated (Cruz-Molina et al., 2017). The results in the current chapter show that RNAPII is present at ~50% of extragenic enhancers identified by other strategies. However many extragenic RNAPII regions are not recovered in published enhancer lists, although they are enriched in enhancer marks, both canonical and non-canonical, and in Polycomb mark H3K27me3. Extragenic RNAPII regions are associated with RNAPII complexes in different states of activation, from poised to fully active. They show diverse enrichment of transcription factors, histone modifications, and proteins involved in chromatin looping, suggesting different mechanisms lead to their formation. For example, Brd4, Med1 and Cohesin are found enriched at increasingly active extragenic RNAPII regions, whereas CTCF, a protein involved in chromatin insulation, is not enriched at any of the extragenic RNAPIIS5p regions, suggesting that extragenic RNAPII and CTCF are not involved in the same processes.

Therefore, extragenic RNAPII identifies candidate regulatory regions and its state relate with their chromatin state.

5.5.2 RNAPII transcription at extragenic regions is regulated similar to genes

RNAPII presence at extragenic regions does not by itself imply a role or a function in gene regulation, and leads to the synthesis of detectable extragenic RNAs, with varying abundance and maturation levels.

Transcription at extragenic regions is regulated at multiple levels. RNAPII kinases play a major role at extragenic regions, as their impairment by chemical inhibition or knockout can drastically reduce RNA transcription and affect RNAPII occupancy. Inhibition of Cdk9 leads to a reduction in RNA transcription at extragenic regions, detected at the level of nascent transcripts, with higher effects on regions occupied by RNAPIIS5p-S7p and RNAPIIS5p-S7p-S2p. Erk2 knock out strongly affects RNAPII occupancy at all extragenic regions leading to a lower occupancy of total RNAPII and RNAPIIS5p, whereas at coding regions it has a specific effect at Polycomb-repressed regions occupied by poised RNAPIIS5p but not at active genes associated with elongating forms of RNAPII (Tee et al., 2014). Sensitivity of RNAPII to perturbation by different transcription regulators suggests that transcription at enhancers may be less stable than at genes, as was proposed for enhancers detected in Drosophila and dependent on Spt5 activity (Henriques et al., 2018).

RNAs transcribed by extragenic RNAPII are under the control of the Exosome machinery, as shown increase RNA detected from extragenic regions after Exosome knock down, and by the measurable occupancy of Exosome subunit Dis3 at the same regions. This is in line with previous reports that showed eRNA sensitivity to Exosome degradation (Andersson et al., 2014a; Arner et al., 2015; Flynn et al., 2011). RNAs were not detected from H3K27me3-RNAPIIS5p regions even after Exosome KD, suggesting that RNAPII may not transcribe these regions to detectable levels. The mechanisms of transcription present at extragenic regulatory regions may have functional role in enhancer activity, while transcripts may in most cases simply be a by-product to be kept under control.

In conclusion, in the current chapter, I used RNAPII to identify extragenic putative regulatory regions and I showed that RNAPII actively produces RNAs with different maturation states, which are concordant with RNAPII state, and that transcription and transcripts are finely regulated.

5.6 Figures Appendix

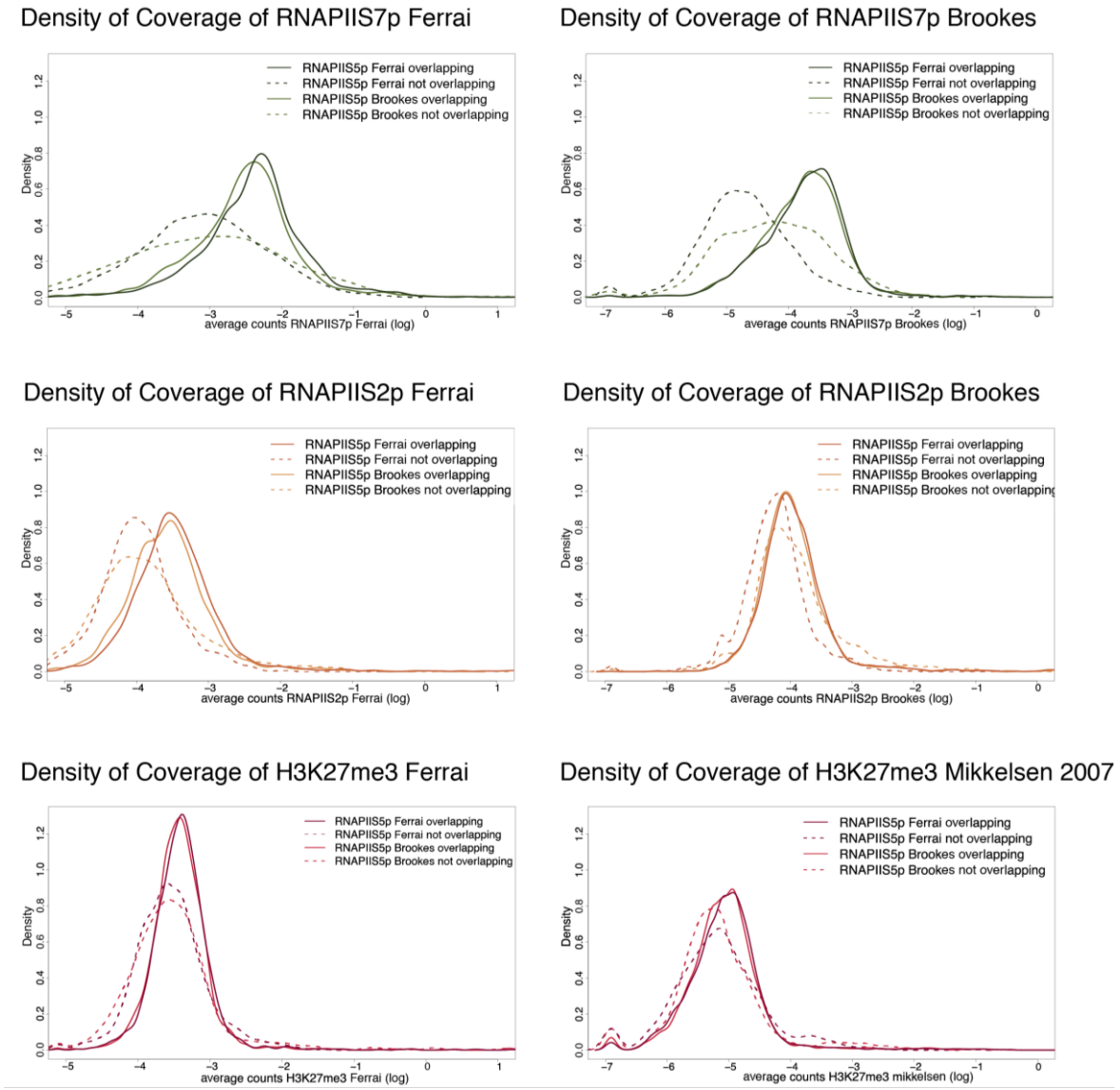


Fig 5.A1: Density distribution of datasets at extragenic RNAPIIS7p, RNAPIIS2p, H3K27me3 regions from Ferrai (dark coloured) and Brookes/Mikkelsen 2007 for H3K27me3 (light coloured). Solid lines represent regions overlapping between datasets; dotted lines represent regions not overlapping within datasets.

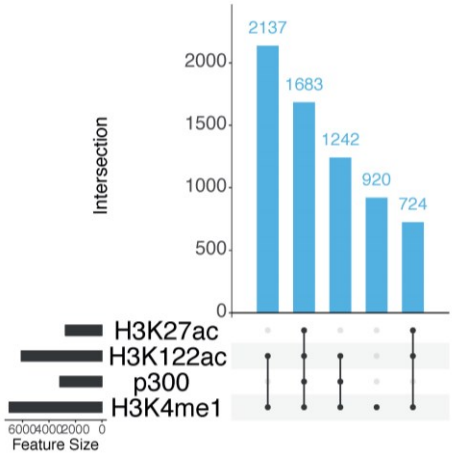


Fig 5.A2: Combinations of enhancer marks at Whyte and Cruz Molina extragenic enhancers.

Figure 5.A3

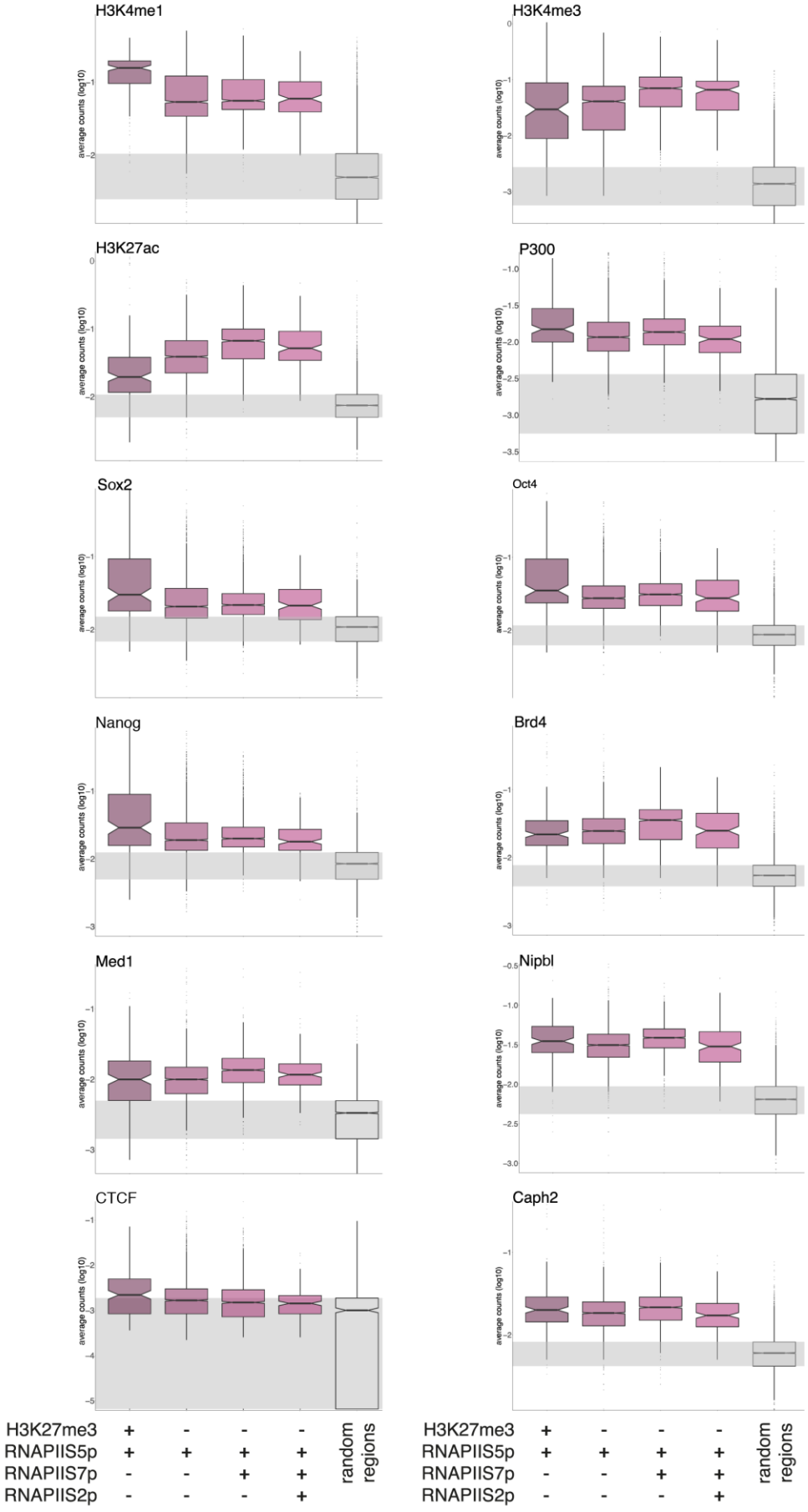

H3K4me1
H3K4me3
H3K27ac
P300
Sox2
Oct4
Nanog
Brd4
Med1
Nipbl
CTCF
Caph2
average counts (log10)
H3K27me3 + - - - random regions
RNAPIIS5p + + + +
RNAPIIS7p - - + +
RNAPIIS2p - - - +

Fig 5.A3: a) Boxplots showing the absolute enrichment of selected features at extragenic RNAPIIS5p regions. Grey area marks enrichment at random regions (25-75% percentile).

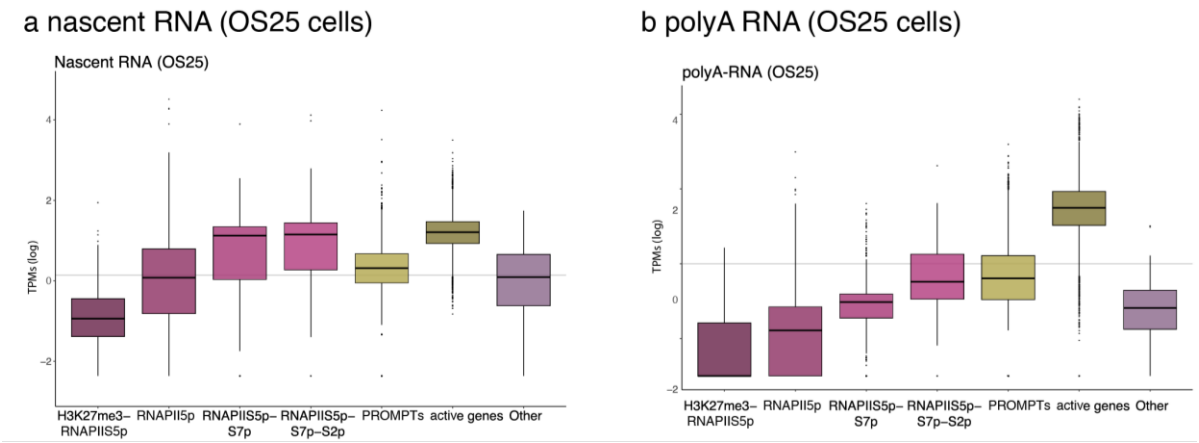


Fig 5.A4: a) Nascent RNA (RNAChIP) from OS25 cells datasets. b) polyA-RNA from OS25 datasets. Log10 TPM + pseudo count are shown for clarity. Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference.

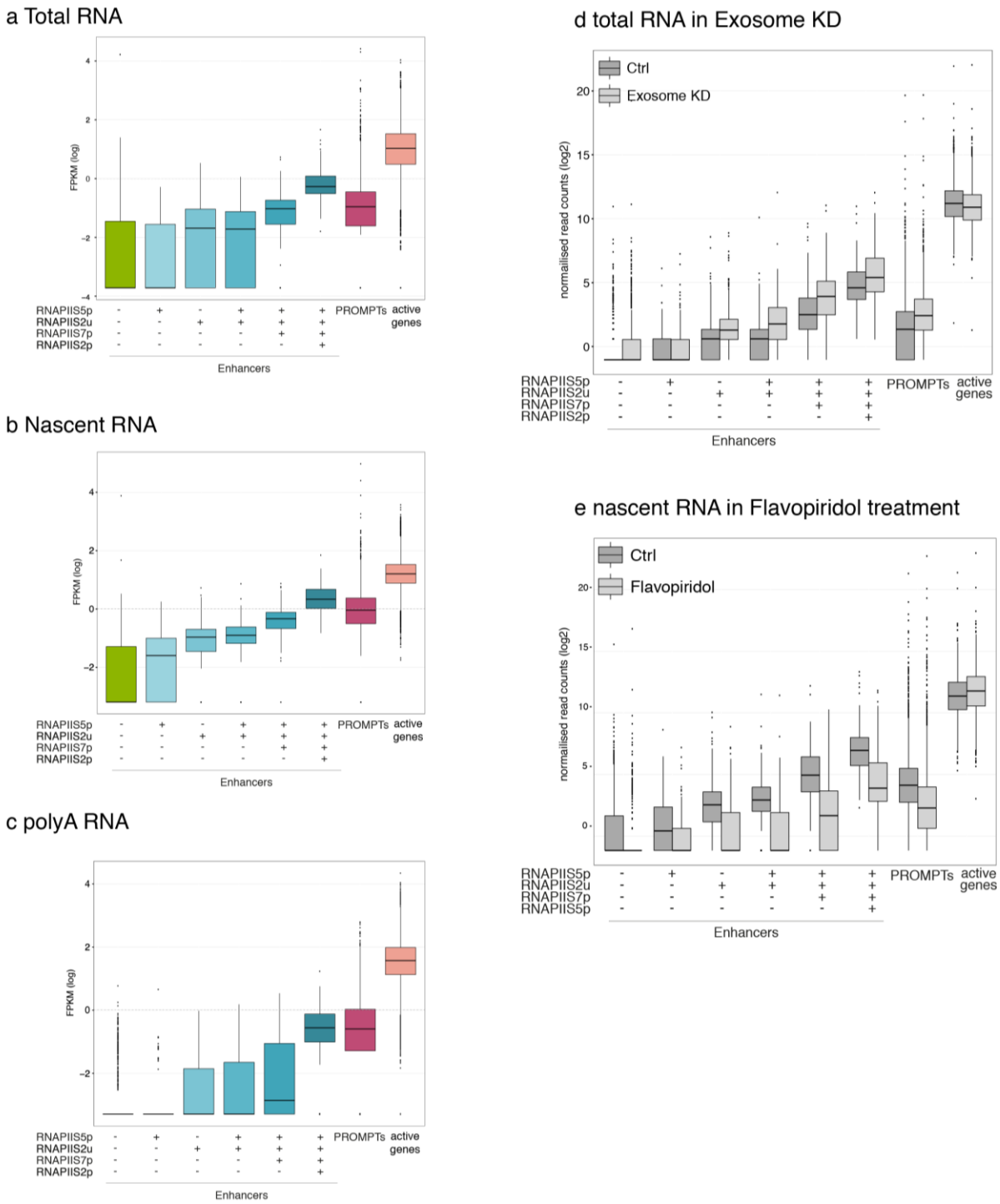


Fig 5.A5: a) total RNA. b) Nascent RNA (RNAChIP). c) polyA-RNA at extragenic Whyte enhancers. Log10 TPM + pseudo count are shown for clarity. Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference. d) Normalised reads of total RNA levels after Exosome KD at extragenic Whyte enhancers. Ctrl: cells treated with a control Si. Si: cell treated with Si against exosome. Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference. e) Normalised reads of nascent RNA levels after Flavopiridol treatment. Ctrl: untreated cells. Si: cell treated with Flavopiridol (1h; 10 uM). Active genes (exon coverage) and PROMPTs regions (500bp) are shown as reference.

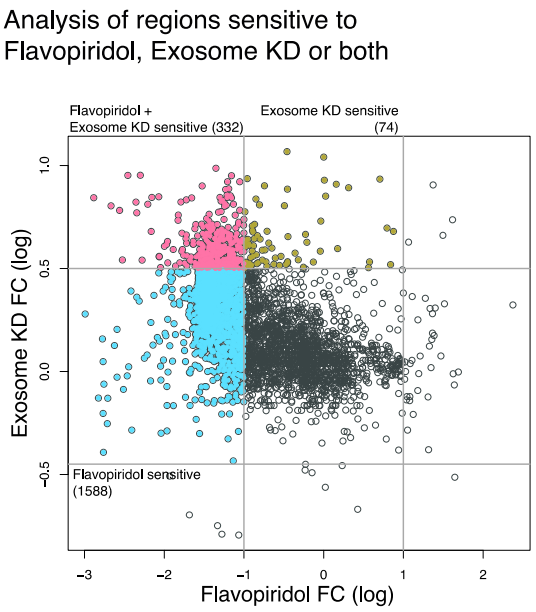


Fig 5.A6: Comparison of sensititvity to Flavopiridol treatment (1h; 10 uM) or Exosome KD. Log of foldchanges calculate with the Deseq2 package in R are shown. Highlighted in blue: regions sensititve to Flaopiridol treatment, in gold: regions sensititve to Exosome KD, in pink: regions sensitive to Flavopiridol treatment and Exosome KD. Every dot represents an extragenic RNAPII region.

6. Extragenic RNAPII identifies active enhancers in neuronal differentiation

The previous chapter dealt with states of activation of extragenic RNAPII in mESC and its relation with enhancer marks and regulation of transcription. In this chapter, I explore whether extragenic RNAPIIS5p regions can function as enhancers *in vivo*.

6.1 Introduction

6.1.1 Validation of putative enhancers

Putative enhancers can be functionally validated using different strategies. Enhancer regions can be cloned in a vector upstream of a weak promoter, and the expression of a reporter gene can be measured as a proxy of enhancer activity (Kvon, 2015). This analysis can be performed at single genes or using high throughput approaches, such as CapSTARR (Vanhille et al., 2015). Recently, a public resource of functionally tested enhancers in mouse and humans has been made available: the VISTA enhancer browser (Visel et al., 2007). This database contains thousands of tested regions in reporter assays in mouse and human (2893 regions tested; 18/4/2018). Specifically in mouse, regions of interest are injected in mouse eggs and tested for reporter activity mainly at the developmental stage E11.5. Regions which give consistent reporter signal (at least 3 positive embryos with a clear pattern) are considered positive enhancer regions. Candidate enhancer regions that are registered as negative in the VISTA database may either not have enhancer function, be active in another developmental time not captured at E11.5, or not have a clear and consistent pattern in at least 3 embryos. Nevertheless, the VISTA enhancer browser is a valuable resource to explore the *in vivo* validity of enhancer predictions.

6.1.2 Neuronal differentiation

To take advantage of the VISTA database for *in vivo* validation of extragenic RNAPIIS5p regions as regulatory enhancer regions, I analysed RNAPII ChIP-seq datasets produced through a neuronal differentiation that captures early stages of development into dopaminergic neurons (Ferrai et al., 2017; Fraser et al., 2015). The neuronal differentiation from Ferrai *et al.* 2017 focuses on 5 time points: mESC (Day 0), early neuronal differentiation (Day1 and 3), and late neuronal differentiation (Day16 and 30). At each stage, cells express markers specific for their

stage (Fig 6.1). At Day 1, cells decrease the expression of pluripotency markers, such as Nanog. At Day 3, cells start to express early neuronal markers, such as Fgf5. At Day 16, cells have a neuronal phenotype, with limited spontaneous firing. At Day 30, cells are post-mitotic and fully developed as dopaminergic neurons. ChIP-seq datasets were publically available for RNAPIIS5p, RNAPIIS7p, and H3K27me3 at all time points. Total RNA-seq datasets were unpublished, but available in the laboratory (A.M. Fernandes, C. Ferrai, unpublished).

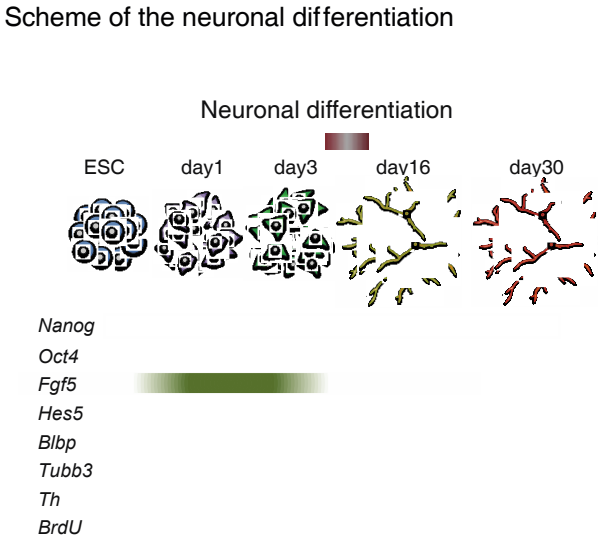


Fig 6.1: Schematic representation of the neuronal differentiation from Ferrai et al. 2017. Scheme representing the time points of the neuronal differentiation and the expression dynamics of cellular marks.

6.2 Aims of the chapter

Enhancer activity is cell, time, and stimulus specific. This is reflected in a high numbers of enhancers described in different cell types (up to 50000 (Andersson et al., 2014a; Nord et al., 2013)) and also in high variability of enhancer usage between different cell types. Understanding how the activity of enhancer changes during differentiation is important to understand their function and their mechanism of action.

To test whether extragenic RNAPII regions identify new genomic regions with enhancer activity during neuronal differentiation, I defined extragenic enhancer regions in each time point of the Ferrai datasets, and classified them according to RNAPII state and Polycomb occupancy (Example of possible dynamic states in Fig 6.2).

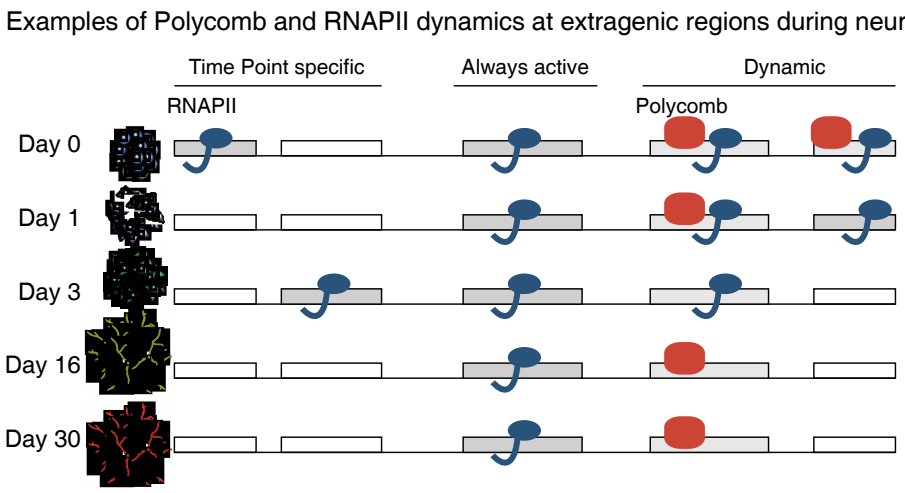


Fig 6.2: Schematic of chapter 6. Schematic representing the possible dynamics of RNAPII and Polycomb during neuronal differentiation

6.3 Contribution disclosure

Elena Torlai Triglia mapped and processed the RNAPII and H3K27me3 datasets and calculated their peaks. Total RNA-seq datasets were produced by Ana Miguel Fernandes and Carmelo Ferrai.

6.4 Results

6.4.1 Datasets used in the current chapter

To study the dynamics of extragenic RNAPII during neuronal differentiation, I took advantage of previously published (Ferrai et al., 2017), and unpublished datasets from the lab (AM Fernandes and C Ferrai, unpublished). For each time point of the neuronal differentiation analysed (Days 0, 1, 3, 16 and 30) were considered ChIP-seq datasets for RNAPIIS5p, RNAPIIS7p and H3K27me3 (Table 6.1) and unpublished total RNA-seq (Table 6.2).

Table 6.1: RNAPII and Polycomb dataset. Table indicating the datasets of RNAPII and H3K27me3 used in the current chapter. RNAPII and H3K27me3 datasets were mapped and processed by Dr Elena Torlai Triglia

Datasets	# Peaks	Time point	Publication	Antibody	GEO
H3K27me3	6677	Day 0 (46c mESC)	Ferrai *et al.*, 2017	anti H3K27Me3, Millipore, # 07-449	GSM2474113
RNAPIIS5p	24010	Day 0 (46c mESC)	Ferrai *et al.*, 2017	anti RNAPII-S5p CTD4H8, BioLegend, # 904001	GSM2474111
RNAPIIS7p	17485	Day 0 (46c mESC)	Ferrai *et al.*, 2017	anti RNAPII-S7p 4E12 (Chapman et al. science 2007)	GSM2474112
H3K27me3	9198	Day 1	Ferrai *et al.*, 2017	anti H3K27Me3, Millipore, # 07-449	GSM2474116
RNAPIIS5p	23446	Day 1	Ferrai *et al.*, 2017	anti RNAPII-S5p CTD4H8, BioLegend, # 904001	GSM2474114
RNAPIIS7p	18051	Day 1	Ferrai *et al.*, 2017	anti RNAPII-S7p 4E12 (Chapman et al. science 2007)	GSM2474115
H3K27me3	9238	Day 3	Ferrai *et al.*, 2017	anti H3K27Me3, Millipore, # 07-449	GSM2474119
RNAPIIS5p	20701	Day 3	Ferrai *et al.*, 2017	anti RNAPII-S5p CTD4H8, BioLegend, # 904001	GSM2474117
RNAPIIS7p	11711	Day 3	Ferrai *et al.*, 2017	anti RNAPII-S7p 4E12 (Chapman et al. science 2007)	GSM2474118
H3K27me3	3810	Day 16	Ferrai *et al.*, 2017	anti H3K27Me3, Millipore, # 07-449	GSM2474124
RNAPIIS5p	34182	Day 16	Ferrai *et al.*, 2017	anti RNAPII-S5p CTD4H8, BioLegend, # 904001	GSM2474120 GSM2474121
RNAPIIS7p	16911	Day 16	Ferrai *et al.*, 2017	anti RNAPII-S7p 4E12 (Chapman et al. science 2007)	GSM2474122 GSM2474123
H3K27me3	5165	Day 30	Ferrai *et al.*, 2017	anti H3K27Me3, Millipore, # 07-449	GSM2474128
RNAPIIS5p	30622	Day 30	Ferrai *et al.*, 2017	anti RNAPII-S5p CTD4H8, BioLegend, # 904001	GSM2474125 GSM2474126
RNAPIIS7p	20268	Day 30	Ferrai *et al.*, 2017	anti RNAPII-S7p 4E12 (Chapman et al. science 2007)	GSM2474127 GSM2474128

Table 6.2: Total RNA-seq datasets. Table indicating the total RNA-seq datasets used in the current chapter. Datasets were produced by AM Fernandes and C Ferrai in the lab

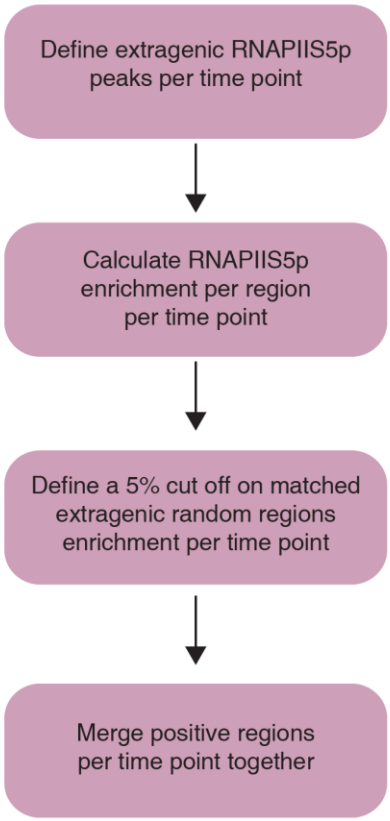
Datasets	Time point	% uniquely mapped reads	Source
Total RNA	Day 0 (46c mESC)	85.68	AM Fernandes, C Ferrai, unpublished
Total RNA	Day 1	83.82	AM Fernandes, C Ferrai, unpublished
Total RNA	Day 3	86.90	AM Fernandes, C Ferrai, unpublished
Total RNA	Day 16	88.27	AM Fernandes, C Ferrai, unpublished
Total RNA	Day 30	87.80	AM Fernandes, C Ferrai, unpublished

6.4.2 Definition of RNAPII extragenic regions across neuronal differentiation

To investigate extragenic RNAPII regions across differentiation, I defined extragenic RNAPII peaks per time point and classified the regions for their enrichment in RNAPIIS5p, as in Chapter 5. Because this analysis was performed during a differentiation timeline and the transcriptional profile of cells varies between time points, I considered as intragenic all the regions overlapping with a RNAPII peak covering a gene at any time point. I then calculate the RNAPIIS5p enrichment at extragenic peaks per time point and the 5% false positive cut-off on matched extragenic random regions, as previously (Chapter 5, section 5.4.2). Extragenic regions identified in this manner in each time point were then merged together to constitute a single list of extragenic regions marked by RNAPIIS5p in at least one time point across neuronal differentiation (Fig 6.3a). RNAPII extragenic peaks at Day 0 were defined in the previous chapter (Chapter 5, section 5.4.4). At Day 1, 93% extragenic RNAPIIS5p peaks were enriched in RNAPIIS5p signal above the 5% calculated threshold; 89% at Day 3; 91% at Day 16; and 92% at Day 30 (Fig 6.3b).

Figure 6.3

a Pipeline to define extragenic RNAPIIS5p regions across time points



b Densities of RNAPIIS5p and threshold

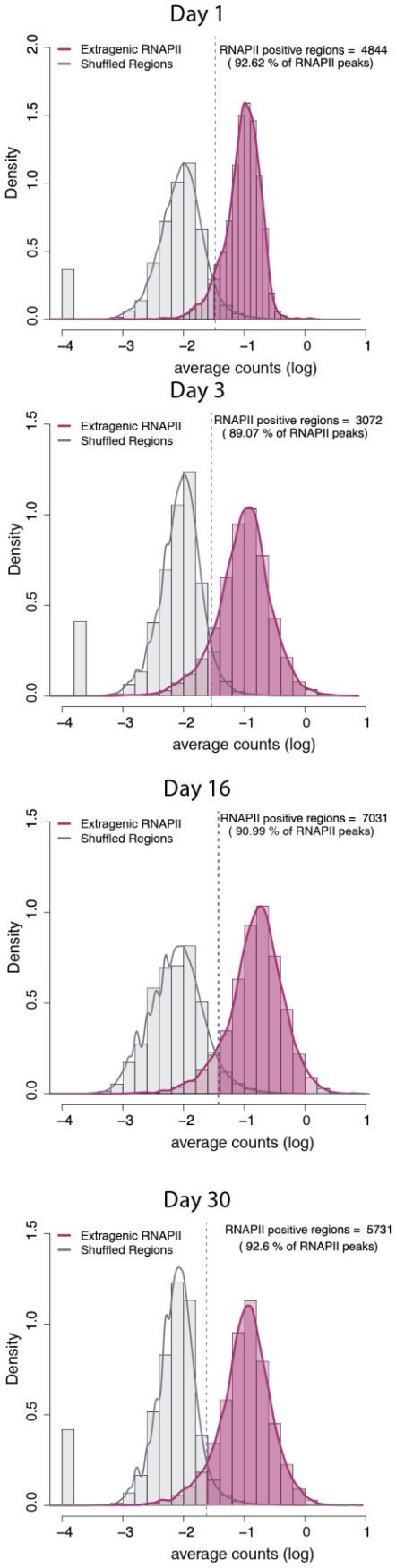


Fig 6.3: Pipeline to define extragenic RNAPIIS5p regions during neuronal differentiation. a) Scheme of the steps of the pipeline to define extragenic RNAPII regions per time point. b) Density plot showing the enrichment per time point of RNAPIIS5p at extragenic RNAPIIS5p regions. Pink thick lines represent maximum enrichment of RNAPIIS5p at RNAPIIS5p extragenic regions per time point. Grey lines represent enrichment at random extragenic regions. Dotted vertical line represent 5% FP cut-off calculated on random regions enrichment.

The full list of RNAPIIS5p regions was then filtered for the ENCODE black list (Consortium, 2012) and merged to obtain a non-redundant list of 16609 extragenic RNAPII regions during neuronal differentiation (Table 6.3).

Table 6.3: Positive extragenic RNAPII. Table indicating the extragenic RNAPII regions before and after applying the 5% false positive threshold per time point and the total of regions after merging and filtering obtain.

	Day0*	Day1*	Day3*	Day16*	Day30*	Total (non redundant)°
Total	5234	5230	3449	7727	6189	-
After threshold	4433	4844	3072	7031	5731	16609

As shown previously, these regions are highly variable between datasets and mainly restricted to one time point (Fig 6.4). With these regions, I moved forward to analyse the extragenic RNAPII dynamics during neuronal differentiation.

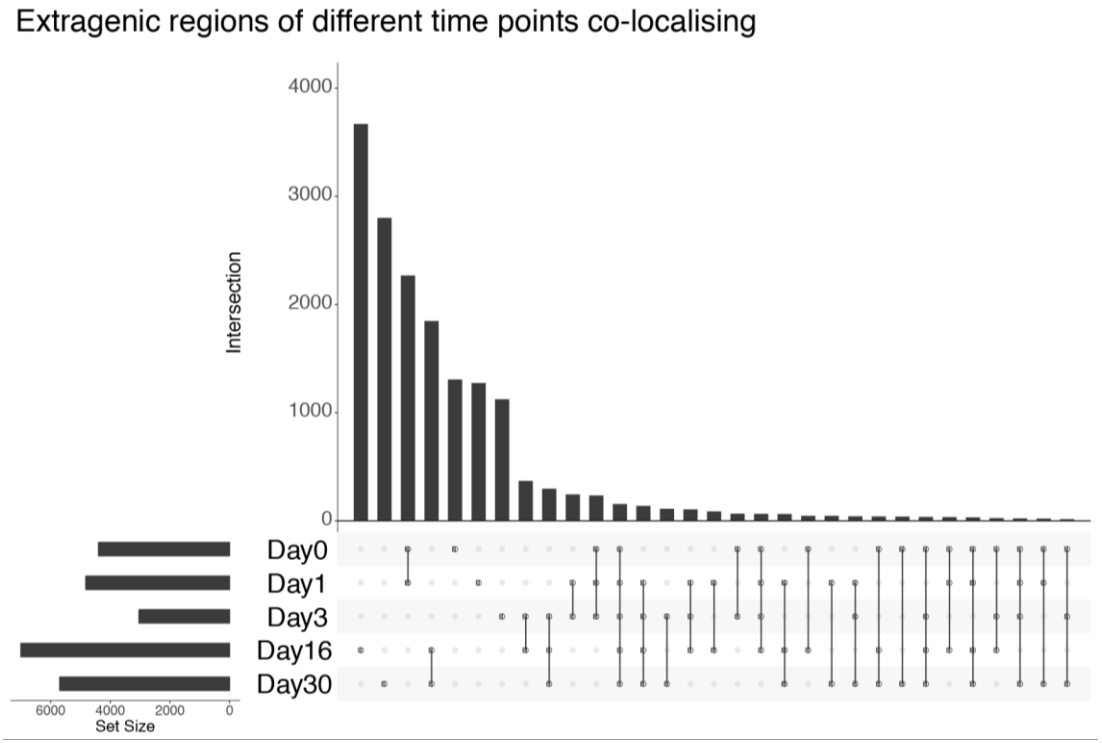


Fig 6.4: Co-localisation of extragenic RNAPII regions across time points. Plot showing the number of combinations of co-localising extragenic RNAPII regions before merging. The majority of regions are unique for a time point or restricted to two.

6.4.3 Classification of extragenic RNAPII states during neuronal differentiation

To analyse the state activation of extragenic RNAPII during neuronal differentiation, I classified the regions according to presence or absence of RNAPIIS5p, RNAPIIS7p and H3K27me3, per each time point (Fig 6.5a), as described in Chapter 5, section 5.4.8. All regions were classified as positive or negative for H3K27me3, RNAPIIS5p, and RNAPIIS7 (Fig 6.5b). The signal of H3K27me3 at extragenic RNAPII regions is very close to the signal of extragenic random regions, leading to a low number of positive regions for this modification, as seen previously for mESCs (Day 0; section 5.4.8).

Figure 6.5

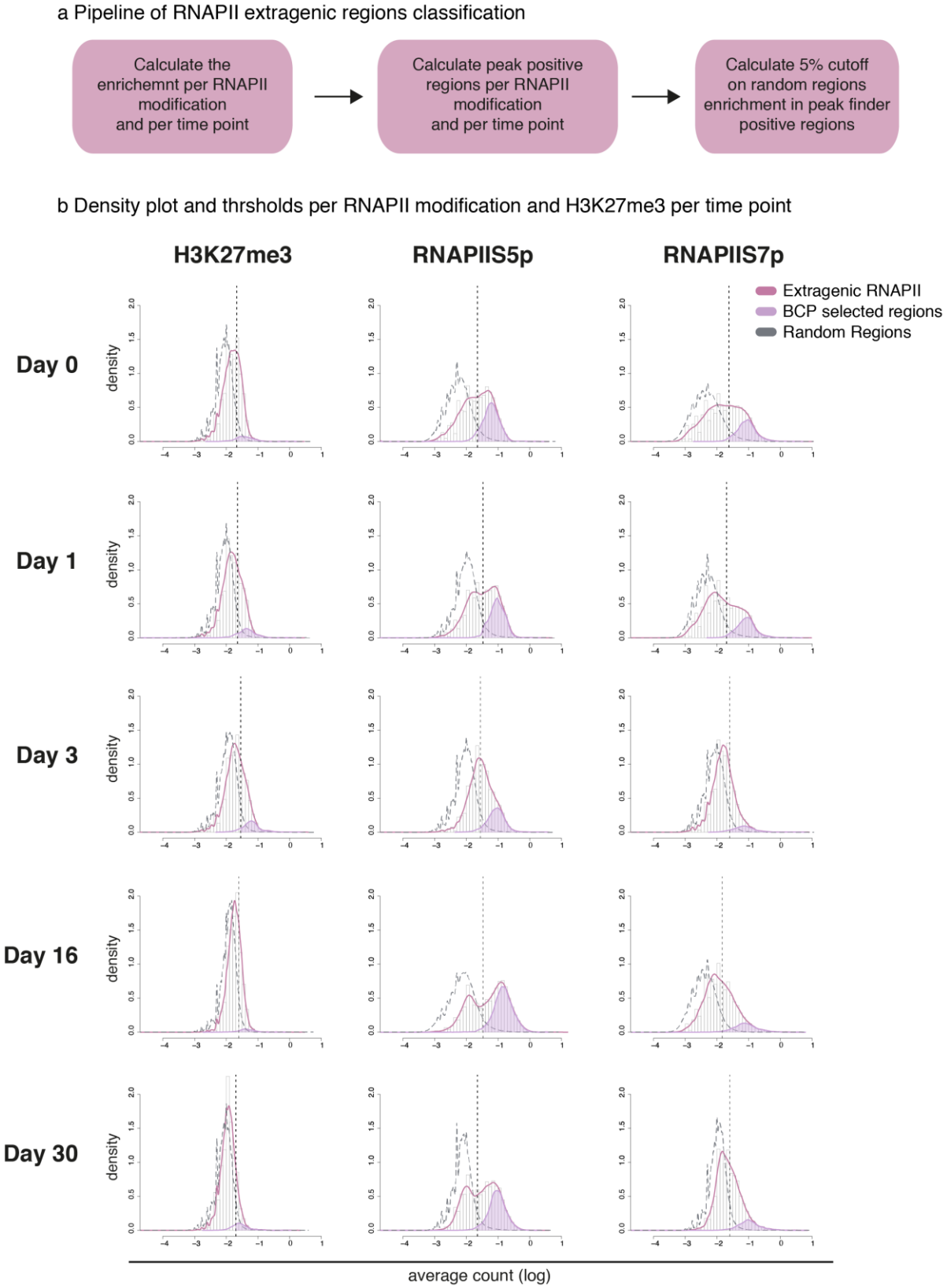


Fig 6.5: Classification of extragenic RNAPII states during neuronal differentiation. a) Pipeline showing the steps taken for extragenic RNAPII states classification per region, feature and time point. b) Classification of H3K27me3, RNAPIIS7p and RNAPIIS5p at RNAPIIS5p extragenic regions. Thin purple lines: enrichment of the considered feature at all extragenic RNAPIIS5p regions; thick purple lines: enrichment of the considered feature at extragenic RNAPIIS5p region overlapping a peak of the featured considered; grey dotted curved line: enrichment at extragenic random regions. Vertical dotted line: 5% FP cut-off.

6.4.4 Extragenic RNAPII is found in different states during neuronal differentiation

RNAPII extragenic regions are found in different states of activation throughout all the neuronal differentiation, which is concordant with previous results in mESC (Fig 6.6).

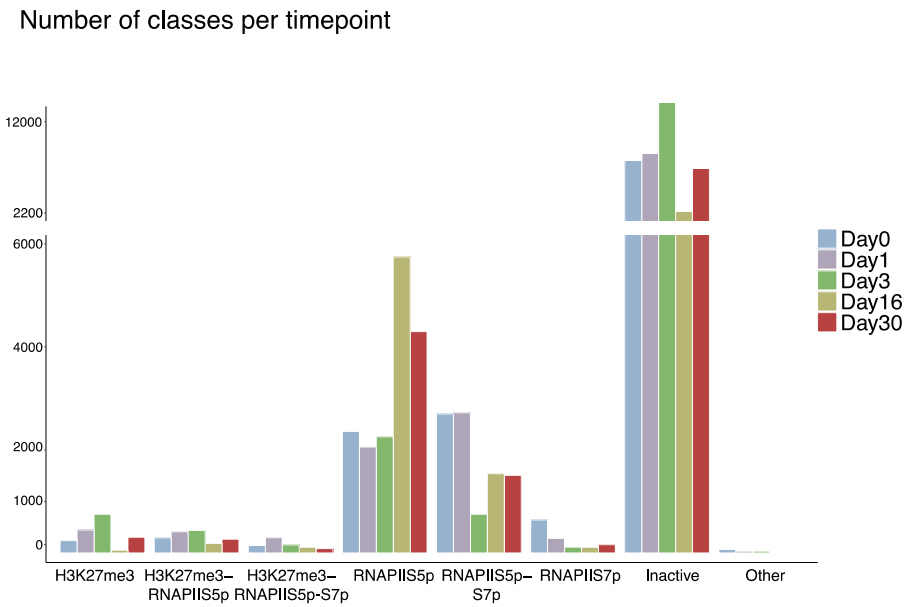


Fig 6.6: Extragenic RNAPII is found in different activation states during neuronal differentiation. Plot showing the number of classes per RNAPII states recovered divided per time point.

In each time point around 30% of the regions are occupied by H3K27me3 and/or RNAPII (Table 6.4). RNAPIIS5p is the most represented class in every time point (30 - 75% of active/Polycomb regions), followed by RNAPIIS5p-S7p (16-42%). H3K27me3 can be found alone (<1-17% of active/Polycomb regions) or together with RNAPIIS5p or RNAPIIS5p-S7p (4-13% of active/Polycomb regions). Classifier enrichment per classes per time point and total RNA transcription can be found in Appendix Fig 6.A1 and show robust classification among time points. Remarkably, H3K27me3-RNAPIIS5p-S7p regions are among the most transcribed classes.

Table 6 4: Extragenic RNAPII states during neuronal differentiation. Table showing the number of extragecni RNAPII activation

	Day0	Day1	Day3	Day16	Day30
H3K27me3	239 (1.4%)	446 (2.6%)	750 (4.5%)	51 (<1%)	302 (1.8%)
H3K27me3-S5p	291 (1.7%)	412 (2.5%)	437 (2.6%)	184 (1%)	267 (1.6%)
H3K27me3-S5p-S7p	144 (<1%)	293 (1.7%)	155 (<1%)	110 (<1%)	84 (<1%)
S5p	2358 (14%)	2055 (12%)	2255 (13.5%)	5747 (34.5%)	4297 (26%)
S5p-S7p	2697 (16%)	2719 (17%)	749 (4.7%)	1537 (9%)	1504 (9%)
S7p	639 (4%)	283 (1.7%)	113 (<1%)	107 (<1%)	160 (<1%)
Other	64 (<1%)	34 (<1%)	34 (<1%)	14 (<1%)	14 (<1%)
Inactive	10177 (61%)	10367 (62%)	12116 (72%)	8940 (54%)	9981 (60%)

In conclusion extragenic RNAPII regions are found in different states of activation during neuronal differentiation. Regions can be repressed by H3K27me3 alone or together with RNAPII, or be in an active state, occupied by RNAPIIS5p and/or RNAPIIS7p. As expected for enhancer regions, a great number of regions are not occupied by either RNAPII or H3K27me3 at each time point, and classified as inactive. Extragenic RNAPIIS5p regions are therefore dynamically regulated.

6.4.5 Extragenic RNAPII regions undergo waves of activation during neuronal differentiation

First, to understand the dynamics of RNAPII and Polycomb at extragenic regions, I analysed the changes in: Active states, with RNAPIIS5p and/or RNAPIIS5p-S7p; Repressed states, with H3K27me3, with or without RNAPIIS5p and/or RNAPIIS5p-S7p; and Inactive states, not occupied by RNAPII and H3K27me3.

Extragenic RNAPII regions dynamically change their activation state during neuronal differentiation. The majority of extragenic regions are active in a single time point or in restricted time frames (Fig 6.7a). Active regions are ~5000 per time point, with a remarkable drop at Day 3, where Polycomb regions are at their highest (1342). Polycomb regions greatly vary between time points with a drop of Polycomb marked regions in differentiated neurons (to 345 in Day 16 and 653 in Day30).

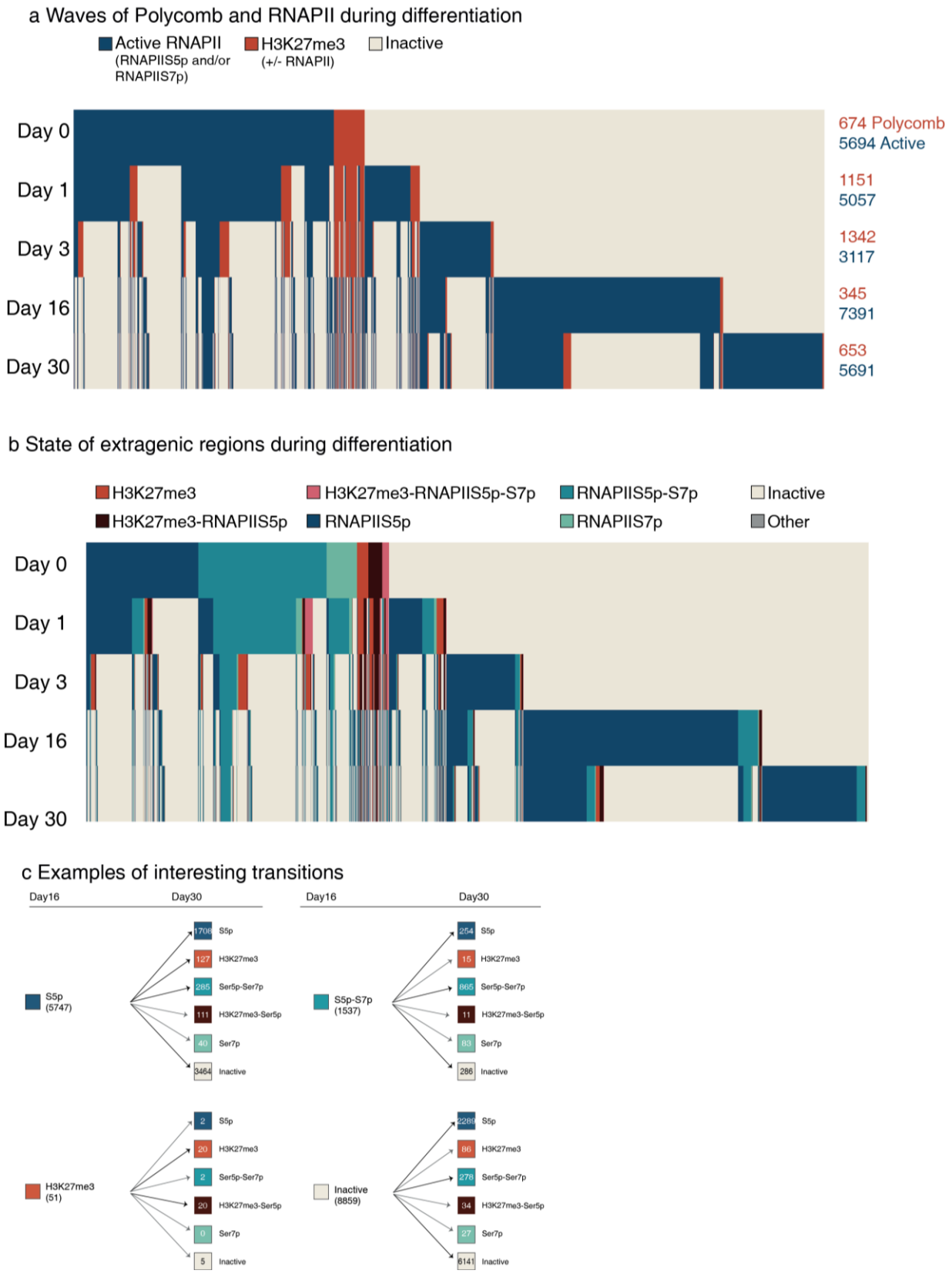


Fig 6.7: Extragenic RNAPII regions dynamic through neuronal differentiation. a) Dynamics of H3K27me3 and RNAPII at extragenic regions during neuronal differentiation. Active regions comprise: RNAPIIS5p, RNAPIIS5p-S7p, RNAPIIS7p and are shown in blue. Polycomb regions comprise: H3K27me3, H3K27me3-RNAPIIS5p, H3K27me3-RNAPIIS5p-S7p and are shown in blue. Inactive regions are shown in cream. b) Dynamics of extragenic RNAPII states during neuronal differentiation. All RNAPII classes specified. c) Schematic representing the most represented transition for selected states and time points from Day 16 to Day 30.

To further understand how RNAPII states differ throughout differentiation, I analysed the transitions all the RNAPII classes (Fig 6.7b). The percentage of transitions of extragenic RNAPIIS5p regions across the differentiation can be found in Appendix Fig 6.A2. The majority of the transitions at active regions keep the RNAPII state or change to Inactive state, while Polycomb repressed regions tend to acquire RNAPII. For example, of the 2358 RNAPIIS5p regions at Day 0, 962 (40%) keep the activation state, 237 (10%) acquire RNAPIIS7p and 953 (40%) become inactive at Day 1 (Fig 6.7c). H3K27me3 regions, on the other hand, preferentially keep their state (Day 0 to Day 1: 54%) or acquire RNAPIIS5p (Day 0 to Day 1: 25%). Inactive regions trough the different time points stay inactive >70% of the time, gain RNAPIIS5p 10-25% of the time, and seldom gain Polycomb (~1%).

In conclusion, these results show that extragenic RNAPII regions are highly dynamic in their state; repressed regions decrease during neuronal development, however they are still present in fully differentiated neurons. This is in contrast with current literature on poised enhancers, which describe the poised state as mainly restricted to embryonic stem cell, as only a minority of poised enhancer were found in differentiated cells (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011; Zentner et al., 2011). RNAPII and Polycomb states show dynamic changes during neuronal differentiation. These regions are time specific and can be either in a repressed or in an active states, which changes through time, resembling the highly specificity of enhancers.

6.4.6 Extragenic RNAPII regions identify enhancer regions active *in vivo*

To investigate whether extragenic RNAPII regions function as enhancers *in vivo*, I analysed the co-localisation of extragenic RNAPII regions with previously tested enhancers of the VISTA enhancer browser (Visel et al., 2007). The VISTA enhancer browser contains ~3000 of tested regions for enhancer activity in mouse embryo at stage E11.5.

To proceed with the comparison, I selected regions from the late development in neurons (Day 16 and Day 30), which should be in part captured in E11.5 (Fig 6.8a). To test whether regions with different states between Day 16 and Day 30 might have different activity in the VISTA database, I selected all the extragenic regions that are active in Day 16 and/or Day 30 (total of ~10000 regions). I compared these regions with regions from the VISTA browser: positive in brain, heart or limb at E11.5; and negative for a specific signal in E11.5. Many (40%) of VISTA regions are negative for specific enhancer signal at E11.5, ~ 30% positive for brain, 20% or less positive for limb and heart (Fig 6.8b). Regions in the VISTA database can be positive for more than one tissue. The analysis with the VISTA enhancer database was conducted with all the

extragenic regions and with the extragenic regions newly identified by extragenic RNAPII and not-colocalising with other published enhancers (Cruz Molina *et al.* 2017, Whyte *et al.* 2013, Chen *et al.* 2012). Strikingly, the results show that extragenic RNAPII regions active are significantly enriched for enhancer activity *in vivo* in brain. Of the newly identified regions: 48% of Active (Day 16) – Active (Day 30) were active in brain, against 32% negative; 48% of Active (Day16) – Inactive (Day30) were active in brain, against 25% negative.

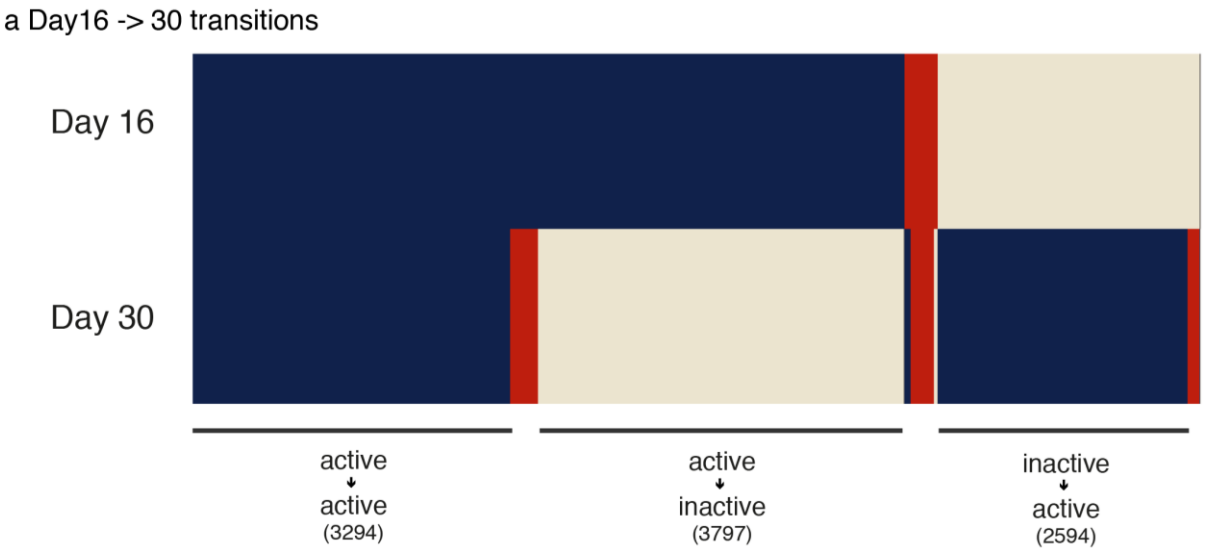


Fig 6 8: Extragenic RNAPII regions identify active enhancers in vivo. a) Dynamics of extragenic RNAPII regions from Day 16 to Day 30. Highlighted are interesting transitions. Active regions comprise: RNAPIIS5p, RNAPIIS5p-S7p, RNAPIIS7p and are shown in blue. Polycomb regions comprise: H3K27me3, H3K27me3-RNAPIIS5p, H3K27me3-RNAPIIS5p-S7p and are shown in blue. Inactive regions are shown in cream. b) Overlap of different transition groups with enhancers from the VISTA database. Bar represent percentage of regions overlapping with VISTA enhancers: in white – negative VISTA enhancers, in grey – positive VISTA enhancers. P-value calculated with hypergeometric test.

Taken together, these results show the power of RNAPII states to mark active enhancer regions *in vivo*. These regions have brain specific activity, in accordance to being defined in neuronal

cells. Similar analysis conducted in the transition between Days 1 and 3 yielded comparable results (Appendix Fig 6.A3).

6.4.7 RNAPII states are linked with enhancer activity *in vivo*

In the previous chapter I showed that RNAPII activity states correlates with enhancer marks enrichment and activation state. To test whether more active RNAPII is more likely to be present at enhancer active *in vivo*, I analysed the state of RNAPII at these regions in more detail. Active enhancers *in vivo* tend to have a more active RNAPII (Fig 6.9a), however to a similar degree to the negative regions. Regions with enhancer activity *in vivo* that undergo inactivation from Day 16 to Day 30 are more transcribed than negative regions (Fig 6.9b), while regions active in both time points are transcribed at similar level regardless of their enhancer activity *in vivo*.

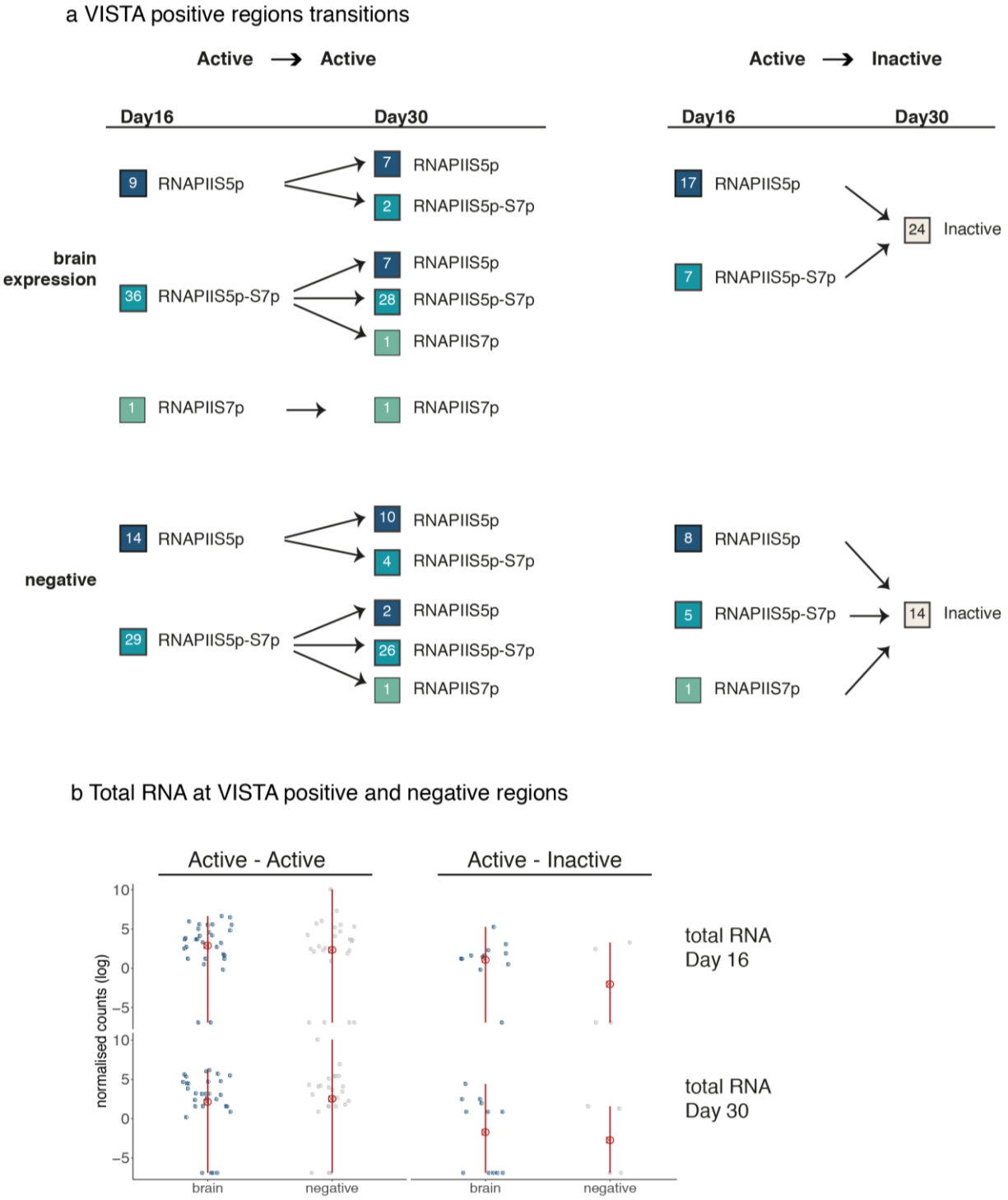


Fig 6.9: Extragenic regions acting as enhancers are in a active state. a) State of extragenic RNAPII regions positive in brain and negative in the VISTA enhancer database. Active to Active and Active to Inactive regions are shown. b) Total RNA in Day 16 and Day 30 of extragenic RNAPII regions positive and negative in the VISTA enhancer database divided by group of transition. Log normailised read counts + 0.001 shown for clarity. Normalised read counts were calculated with the DESeq package in R. Every dot represent a region.

Taken together, these results show that extragenic RNAPII regions co-localising with VISTA regions are bound by RNAPII in an active state, which is likely to transcribe through these regions. Future analyses will explore available RNA-ChIP-seq datasets produced for some of the data points considered (Robert A. Beagrie, Carmelo Ferrai, unpublished).

6.5 Discussion

In the current chapter, I show that extragenic RNAPII regions are highly dynamic during differentiation, and have enhancer activity *in vivo*.

6.5.1 Extragenic RNAPII states dynamics during neuronal differentiation

Extragenic RNAPII is present in different activation states through neuronal differentiation, concordant with previous results in mESC (Chapter 4-5). The dynamics of RNAPII and Polycomb at extragenic regions are more variable than the ones previously seen at promoters in the same neuronal differentiation (Ferrai et al., 2017). Enhancers are more variable in their activity than genes (Nord et al., 2013), which supports the finding that of all the extragenic regions characterised during neuronal differentiation 50-70% are inactive per time point. Noticeably, closer time points share more regions than time points further apart, which is in line with enhancer specific activity. Same extragenic regions in related cells can regulate cell identity genes, which need to be activated. Cells in very different differentiation states, on the other hand, should have fewer expressed genes in common, and possibly fewer shared regulatory regions. In line with this, active regions in one time point tend to either remain active or to turn to inactive and few regions are active at early time points, become inactive, and turn active again. Interestingly, Polycomb regions seldom turn to inactive, while more often either remain repressed or acquire RNAPII. It is possible to speculate that these regions are kept poised and ready to be activated. More studies are needed to understand their target genes and their function.

Some extragenic regions remain active throughout neuronal differentiation. These regions are mainly bound by RNAPII without Polycomb (data not shown) and could constitute enhancers recently described to regulate housekeeping genes (Arnold et al., 2013). While an investigation of these regions was not conducted for time constrains, it would be of interest to analyse these regions further, especially their enrichment for binding sites of expressed TFs, or TF binding profiles, if available, and their involvement in chromatin architecture.

6.5.2 Extragenic RNAPII marks enhancers active *in vivo*

To assess the power of extragenic RNAPII as an approach to identify regulatory regions, I compared these regions with the regions of the VISTA enhancer browser. The VISTA enhancer browser contains thousands of regions tested for their reporter activity in mouse embryos (Visel et al., 2007). The majority of regions were tested at E11.5, and are the ones used for comparison in the current chapter. Remarkably, regions identified with extragenic RNAPII are significantly

enriched in enhancer with specific activity in brain. Regions positive for enhancer activity *in vivo* have an active transcribing RNAPII. This reflects the findings of chapters 4 and 5, where I showed that RNAPII state are indicative of enhancer state. These results were obtained for regions active in Days 1 and 3 (early development) and Days 16 and 30 (late development) and highlights that extragenic RNAPII can specifically identify regions with enhancer activity *in vivo*. Regions active specifically in Day 30 and co-localising with VISTA enhancers were few and mainly negative, as expected as Day 30 neurons have mature electrophysiology responses and fully developed dopaminergic neurons develop between E11 and E14, which can explain why they are not captured in the comparison with VISTA regions (Ang, 2006). Extragenic RNAPII was able to recover specifically regions active in brain development. It would be of interest to apply this method to heart development, for example, as it is also covered in the VISTA database.

6.6 Figures Appendix

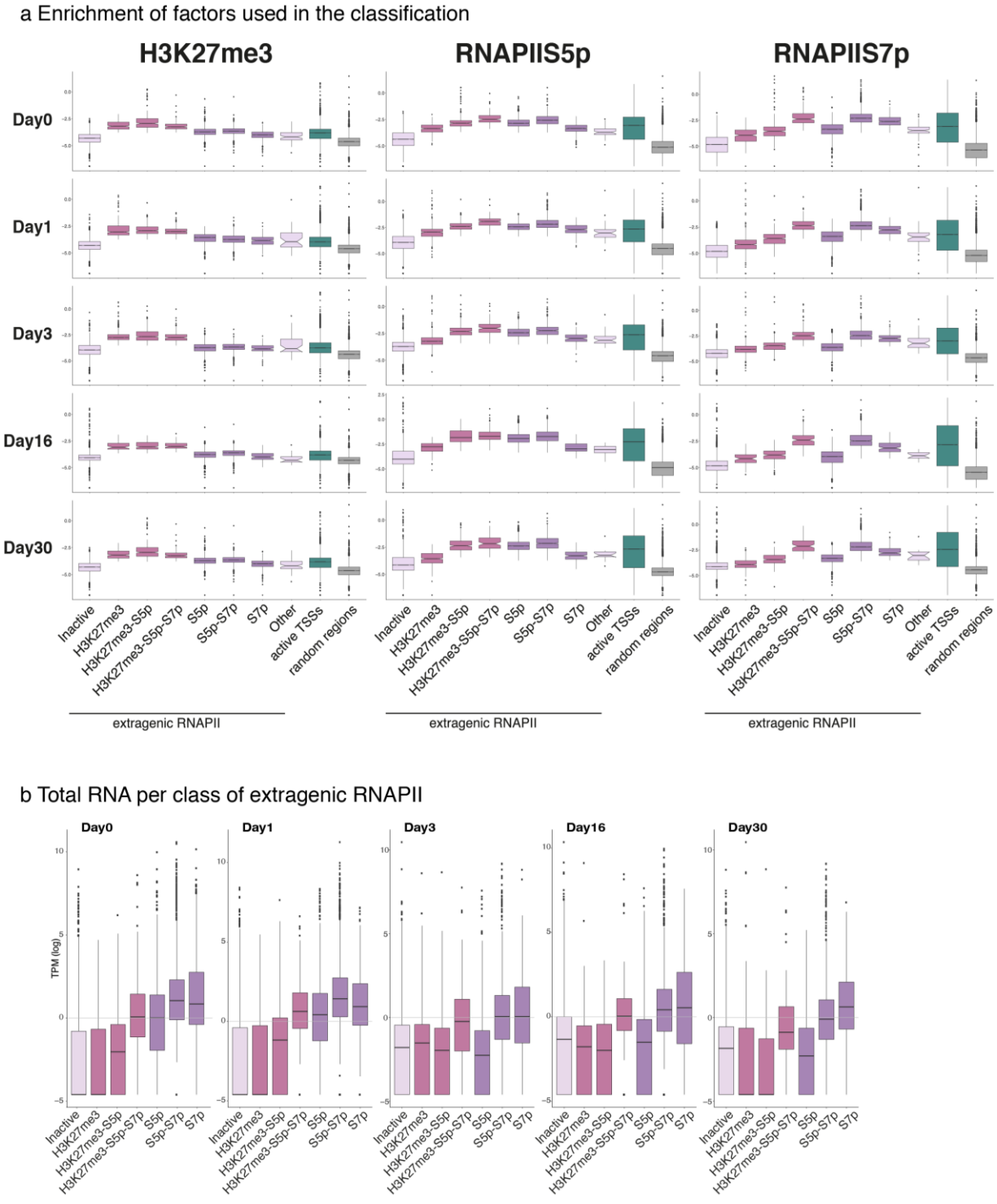


Fig 6.A1: Extragenic RNAPII activation states across differentiation are robust. a) Boxplots showing the enrichment of classifiers per time point per class. Active TSSs in the specific time point and extragenic random regions are shown as reference. Active TSSs were downloaded from Ferrai et al. 2017 and are +/- 2kb from the active TSS. Boxplots show the log average count of ChIp-seq reads per region + pseudo count are shown for clarity. b) Total RNA transcribed per class of extragenic RNAPII per time point. Log TPM + 0.001 shown for clarity.

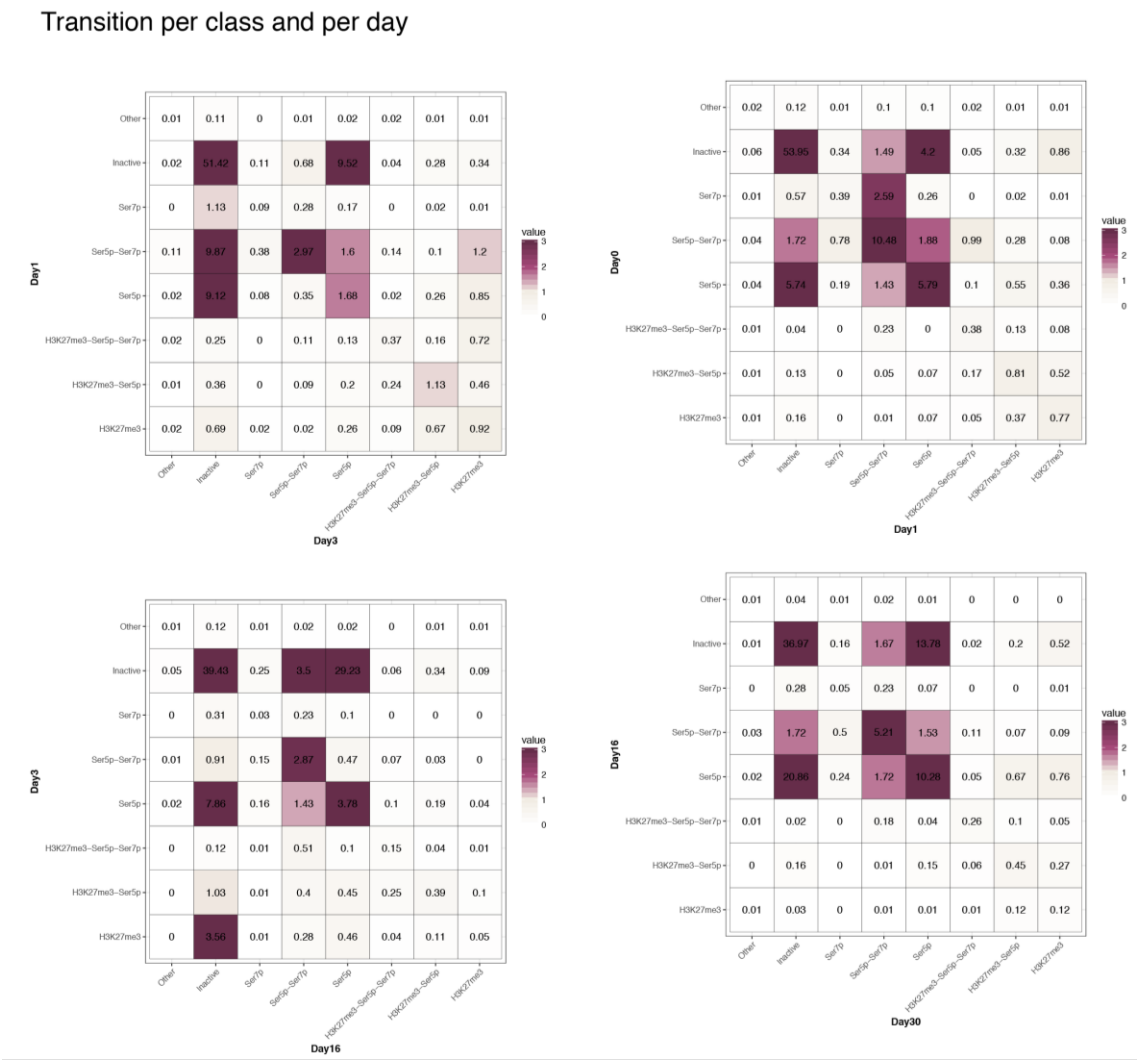


Fig 6.A2: Transition of all extragenic RNAPII states. Plots showing the total percentage of the transition per day pair of extragenic RNAPII states.

Figure 6.A3

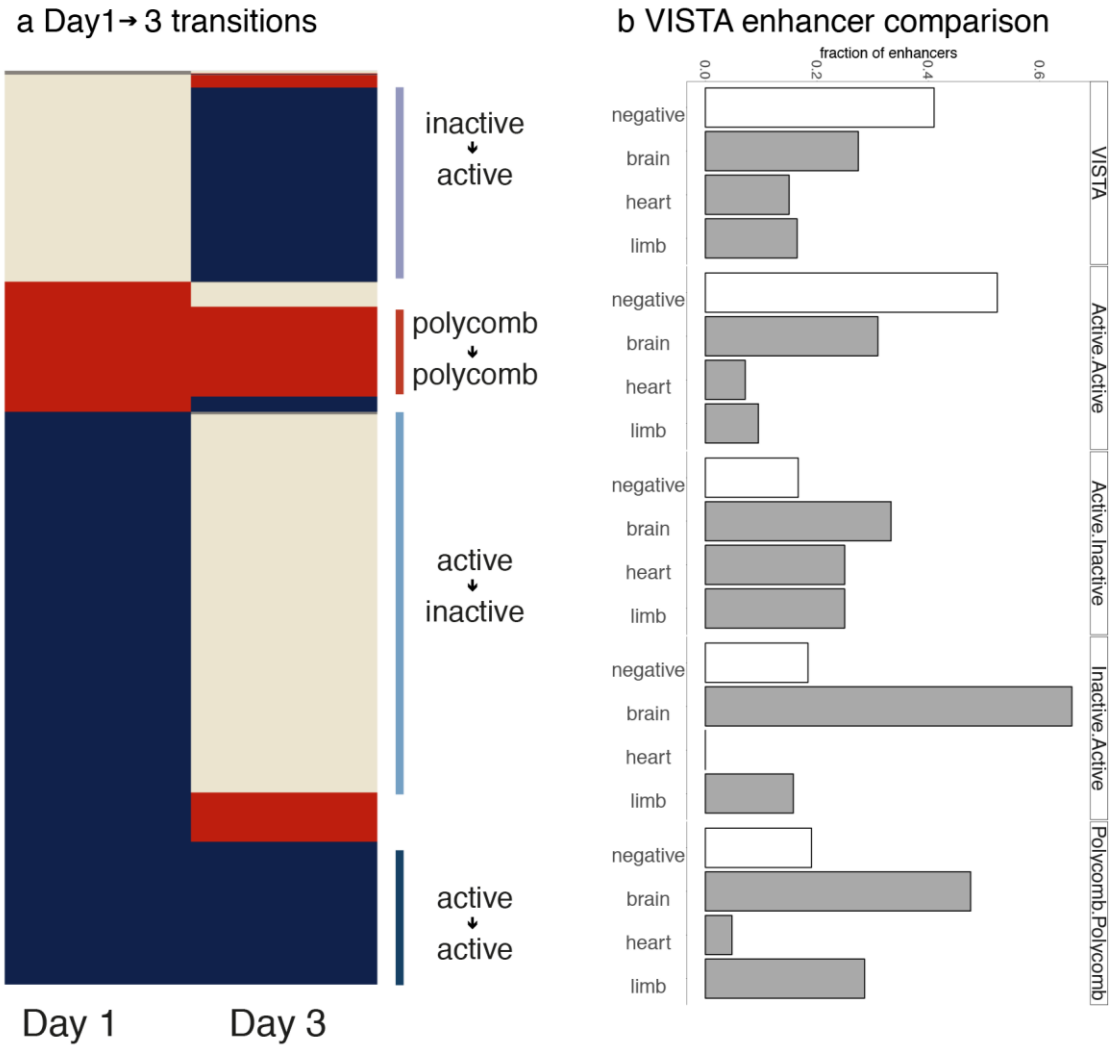


Fig 6.A3: extragenic RNAPII dynamics in early differentiation identify enhancer regions involved in development. Dynamics of extragenic RNAPII regions from Day 1 to Day 3. Highlighted are interesting transitions. Active regions comprise: RNAPIIS5p, RNAPIIS5p-S7p, RNAPIIS7p and are shown in blue. Polycomb regions comprise: H3K27me3, H3K27me3-RNAPIIS5p, H3K27me3-RNAPIIS5p-S7p and are shown in blue. Inactive regions are shown in cream. b) Overlap of different transition groups with enhancers from the VISTA database. Bar represent percentage of regions overlapping with VISTA enhancers: in white – negative VISTA enhancers, in grey – positive VISTA enhancers.

7. Discussion

The work presented in the current thesis is part of a fast-moving field that tries to characterise regulatory regions and understand their features, and at the same time aims to find new, powerful approaches to identify regulatory regions. This is particularly important to discover target genes affected by disease-associated sequence variation in non-coding regions of the genome, which are likely to have regulatory or enhancer functions. Enhancers are described as regions outside promoters, which regulate target gene expression. However, it is not clear what is the difference between enhancers identified using different strategies and which features are most powerful to identify and characterise enhancer activation states.

Enhancer identification approaches

In Chapter 3, I compared the genomic position and other features of candidate enhancer regions identified genome-wide using different approaches. Each approach uses different chromatin occupancy features, which leads to differences in the type of regions identified. However, the size and implications of this effect have been often overlooked. Not only different regions are detected in different approaches, but perhaps more difficult, regions classified in different activation states can co-localise, such as super enhancers and poised enhancers. Results in chapter 3 show that similar approaches can identify enhancers with different characteristic, probably for technical reasons such as differences in datasets, peak calling methods or the thresholds used. An important result of my research was the finding that regions enriched for enhancer marks and transcription factors are also bound by RNAPII, independently of the approach used for the identification.

Different states of activation of RNAPII at extragenic enhancer regions

While different studies analysed RNAPII occupancy at extragenic regions and remarkably some also used different RNAPII modifications datasets(Cruz-Molina et al., 2017; De Santa et al., 2010; Estarás et al., 2015; Koch et al., 2011), the state of RNAPII at enhancers was not previously investigated in depth. In Chapter 4, I showed that RNAPII at extragenic enhancers exists in different states of activation from transcription initiation, marked by RNAPIIS2u and/or RNAPIIS5p, to elongation, marked by RNAPIIS2p. The activation state of RNAPII mirrors the enhancer activation state, which suggests that RNAPII states can be used as readout of enhancer states and for enhancer identification in the genome.

Identify of candidate regulatory regions using RNAPII occupancy in mESC

In Chapter 5, I produced a list of candidate enhancers identified based on RNAPIIS5p occupancy at extragenic regions and showed that almost all of them overlap with enhancer marks: only 30% of them had been previously catalogued as enhancers with other approaches. Remarkably, extragenic RNAPII regions show the same correlation between the state of RNAPII and the enrichment for enhancer marks, irrespectively of whether they overlay with previously identified candidate enhancers. RNAPII occupancy is therefore a valuable approach to identify regulatory regions. Moreover, the observation that RNAPII at extragenic regions transcribes RNAs with different states of maturation, some with properties of eRNAs, confirms the occupancy of RNAPII at these extragenic regions and shows that the RNAPII state reflects the maturation state of the RNA. Transcription at extragenic regions is regulated through modifications of the RNAPII-CTD that are important for promoter pausing release mechanism.

RNAPII regions in a neuronal differentiation time course

In Chapter 6, to test the power of RNAPII to identify enhancers, I detected extragenic RNAPII regions in a neuronal differentiation time course, and showed that these regions undergo extensive dynamic states between time points, from inactive, to Polycomb repressed, to active. Importantly, the candidate regions identified using in vitro differentiated neurons with extragenic RNAPII show enhancer activity *in vivo* specific for the brain, both for candidate regulatory regions newly identified or previously identified by other approaches.

Differences between enhancer identification approaches: what we can learn

Enhancers are known since 1981 and were first described in SV40-infected HeLa cells as a genomic element capable of activating a target gene independently of the orientation or the position in respect to the target gene(Banerji et al., 1981). Since then, various approaches were developed to identify and characterise enhancers. Putative enhancers can be defined in different states of activation, which are indicative of their functions: for example, active enhancers will enhance, while poised enhancer repress target gene expression(Buecker et al., 2014; Calo and Wysocka, 2013). Approaches to identify enhancer regions can vary in their methodology, leaving open the question of which has a higher sensitivity and specificity. An interesting study (Benton et al., 2017) aimed to understand if enhancers regions identified with more than one approach have a tendency to be more active. The authors showed that not only this was not the case, but also that different approaches found regions with different characteristics, such as

conservation, GC content, etc. Benton and colleagues (2017) focused on active enhancers. In the current work, I compared not only enhancer lists defined with different approaches, but also enhancers classified in different states of activation.

The major conclusion that emerges from the comparisons between enhancer lists is that each enhancer detection approach has its specific biases depending on the chromatin occupancy features used as input. These biases should be taken into account depending on which type of enhancer one aims to discover. For example, candidate enhancer regions bound by TFs can be actively transcribed by RNAPII, and enhancers occupied by histone modifications that mark active enhancers, such as H3K27ac, can be bound by proteins involved in chromatin looping, such as CTCF (Hansen et al., 2016). Enhancer regions identified by different criteria may exert different functions. As suggested by Benton and colleagues (2017), it is possible that enhancer regions as currently defined are an aggregate of different types of regions, which may regulate through diverse mechanisms. For example, an interesting finding emerging from the current work is the diverse relation between specific transcription factors and RNAPII at enhancers. E2f1 enrichment showed a high correlation with RNAPII binding, whereas Smad1 enrichment is inversely correlated, in line to repressive roles of the Smad family in gene regulation(Vincent et al., 2009). E2f1 was shown to regulate Brd4, an important factor bound at super enhancers(Ma et al., 2003). These observations suggest that regulatory regions can exert different functions, repressive or activating, based on their respective marks, which are complex and yet to be fully understood.

In the future more studies on the relation between regulatory regions and their functions, such as involvement in direct contacts or in insulator mechanisms, as well as dynamics during cell differentiation or after stimuli, will help clarify the diversity of regulatory regions and potentially distinguish between regions with diverse function.

RNAPII state as readout for enhancer state

Enhancers were described to exist in different states of activation (Calo and Wysocka, 2013) and to transcribe enhancer RNAs (eRNAs). eRNA transcription is linked with enhancer activity (Kim et al., 2010;Kaikkonen et al., 2013), and it has been used to identify enhancers in the genome (Arner et al., 2015; Henriques et al., 2018). I decided to take a complementary approach and define enhancers based on extragenic RNAPII occupancy, with the aim of understand if it is possible to define the state of enhancers and genes with a minimal dataset. To conduct analyse RNAPII states at enhancer regions, I focused on extragenic regions, to avoid the confounding effect of gene-coding transcription. Though enhancers are known to also reside inside gene

introns, these regions were not considered in this work. A future possibility to study RNAPII at intragenic regions would be to consider enhancers inside inactive genes; inactive genes would be transcriptionally silent, so it would be possible to identify RNAPII states at intronic or exonic enhancers of inactive genes.

Enhancers bound by RNAPII show features associated with active and poised enhancers, such as variable levels of transcription and typical chromatin marks, e.g. H3K27ac, H3K4me1, TFs and P300. Interestingly, TF binding alone is not able to distinguish between differently active enhancers, in line with what was previously shown for super enhancers and normal enhancers (Hnisz et al., 2013; Whyte et al., 2013).

RNAPII states at enhancers recapitulate the states found at genes. This finding clarifies an open question in the field about whether or not RNAPII at enhancers shows differences from RNAPII at genes. Although previous work failed to identify RNAPIIS2p at enhancer regions (Koch et al., 2011; Li et al., 2016), the work presented in this thesis clearly shows that a subset of enhancer regions are bound by RNAPIIS2p and transcribe detectable polyA-RNA. Low levels of polyA-RNAs originating from enhancer regions were previously described (De Santa et al., 2010). Although RNAPIIS2p-positive enhancers are the most transcribed enhancers in nascent and total RNA-seq datasets, the level of expression is considerably lower than that of active genes. Moreover, RNAPIIS2p enhancers, together with the other RNAPII-bound enhancers, are sensitive to transcription perturbation using Cdk9 inhibition, Erk2 knockout, and Exosome knockdown, the latter contrary to active coding genes.

One interesting observation was the identification of candidate enhancer regions occupied by RNAPIIS2u alone. RNAPII at genes is phosphorylated on Ser5p upon binding (Brookes and Pombo, 2009). Detection of RNAPIIS2u alone is uncommon at genes, while it is a numerous class at extragenic regions. It is possible that phosphorylation at enhancers is less efficient, maybe because of the lower abundance of kinases such as Cdk8 and Cdk7 at extragenic regions compared to promoters (data not shown). In line with this theory are the results that show that RNAPIIS2u at extragenic regions has intermediate enrichments of active features and low transcription levels. However, it is also possible that RNAPII at these regions is present in a different form not considered in the current work. For example, RNAPIIS2u could be in an initiating stage marked by RNAPIIK7me1/2, modifications of the distal non-consensus CTD that was linked with early stages of transcription at active genes (Dias et al., 2015). Less probable would be the presence of Tyr1p, which was associated with elongating RNAPII (Mayer et al., 2012a).

Finally, it remains to be understood whether regions bound by TFs, but depleted of RNAPII, have a different enhancer state or act with different mechanisms. For example, it is possible that enhancer regions without RNAPII would be involved in longer chromatin loops (Rao et al., 2014) which define the preferential space of activity of other enhancers in the chromatin locus. Topological Associated Domains (TADs) are described as regions of preferred interaction (Dixon et al., 2016), and at least some TAD borders which are enriched in CTCF act as insulators. Loops and TADs, however they are defined, are similar in concept, where structural proteins confine large stretches of chromatin to interact between themselves. TADs and loops were found to be to some degree cell specific(Fraser et al., 2015; Rao et al., 2014). In this frame, enhancers with and without RNAPII can be different genomic elements that work together to regulate gene expression. For example, enhancers without RNAPII could have a structural role, while enhancers with RNAPII could contact the genes and create regions enriched for TFs and co-factors which can promote transcription (Beagrie and Pombo, 2016; Merika and Thanos, 2001). It has to be noted that TAD boundary can be enriched in RNAPII (Fraser et al., 2015) and this could also suggest that TAD boundaries have different natures.

Extragenic RNAPII as a powerful approach for enhancer identification

Identification of putative enhancers remains a major challenge in the field. In the current work, I present a novel way to identify enhancers using RNAPII occupancy. Extragenic regulatory regions identified with this approach overlap >95% of the time with known enhancer marks and show typical enhancer features, such as Exosome sensitive transcription and high dynamic states during differentiation. Previously, De Santa and colleagues showed that RNAPII marks putative enhancers regions in macrophages (De Santa et al., 2010), however the work presented in this thesis substantially advance their initial findings, demonstrating that different states of activation of RNAPII distinguish differently active enhancers. Furthermore, I demonstrate that extragenic RNAPII regions are able to find enhancer regions active *in vivo* in specific cell types. The use of extragenic RNAPII to identify enhancers proved to be powerful also in identifying Polycomb repressed enhancers, that would be missed by methods based on transcription, due to the low to not detectable levels of eRNAs originating from these regions (this work and Rada-Iglesias *et al.*(Rada-Iglesias et al., 2011), Cruz Molina *et al.(Cruz-Molina et al., 2017)*). Another advantage of using RNAPII to identify regulatory regions is the possibility, with a minimal dataset, to be able to evaluate the activation state of both promoters and enhancers of a cell. RNAPIIS5p, RNAPIIS7p, and H3K27me3, plus RNA-seq were proven to be sufficient to characterise promoter states in mESC and during differentiation (Brookes et al., 2012; Ferrai et al., 2017).

This works adds the opportunity to identify and characterise the enhancers with the same datasets. Not only would it be possible to have promoter and enhancer states at the same time, but even more importantly RNAPII datasets do not need previous knowledge on TFs important in the specific cell type or reaction to stimuli, reducing a potentially issue to experimental design. Potentially, RNAPII extragenic regions specific for a developmental stage or a stimuli response could be used to infer TF binding via transcription factor binding motif algorithms. For example, it would be interesting to understand which TFs bind enhancers bound by RNAPII at early and late stages of neuronal differentiation Interestingly, some extragenic RNAPII regions identified in this work during neuronal differentiation are active across all the time points (data not shown). It would be of interest to understand the specificity of these regions, for example if they regulate housekeeping genes. Housekeeping genes were recently shown to be regulated by constitutive enhancers (Zabidi et al., 2014). Extragenic RNAPII could potentially be able to identify also these regions, together with poised enhancers and time-specific enhancers.

In the future, it would be of great interest to understand whether specific RNAPII activation states at enhancers are indicative of target gene promoter states, and whether this information could be used to identify target genes. For example, poised enhancers can regulate repressed genes and the specificity of this regulation could be mediated by RNAPII. RNAPII states at enhancers and promoters could be used to infer target genes. Although this idea needs to be proven, it would potentially lead to the possibility of characterising enhancer states, promoter states, and enhancer promoter contacts with a minimal datasets featuring RNAPIIS5, RNAPIIS7p, and Polycomb.

Transcription at extragenic RNAPII identified enhancer is regulated similarly to genes

Enhancers were found to transcribe RNAs (Andersson et al., 2014a; Arner et al., 2015; Kim et al., 2010;Kaikkonen et al., 2013). eRNAs are detected with techniques such as CAGE (Andersson et al., 2014a; Arner et al., 2015), total RNA (Koch et al., 2011), GRO-seq (Danko et al., 2015), Start-seq (Henriques et al., 2018), which give indications of the nature of these transcripts. My analysis showed that nascent RNA has the highest signal in the different extragenic RNAPII regions classified. eRNAs were shown to be transcribed and degraded by the Exosome machinery (Andersson et al., 2014a), which happens also at RNAPII extragenic regions. Moreover, extragenic regions are sensitive to Flavopiridol, a drug that inhibits Cdk9. Cdk9 inhibition block productive elongation because it interferes with RNAPIICTD phosphorylation and promoter proximal pausing release. Recently, Henriques and colleagues

characterised the role of the pausing factor Spt5 in transcriptionally active enhancers in Drosophila (Henriques et al., 2018), which is in line not only with the finding on Flavopiridol sensitivity, but also with the observation that Spt5 and Nelfa bind extragenic RNAPII regions. Interestingly, RNAPIIS7p and RNAPIIS2p presence correlates with Spt5 and Nelfa binding, with Nelfa more enriched at RNAPIIS7p regions without RNAPIIS2p and Spt5 enriched in both. Spt5 travels with elongating RNAPII while Nelfa detaches after promoter proximal pausing release. This observation would suggest that RNAPIIS5p-S7p-S2p extragenic regions would be in a more active and elongating phase of transcription compared with the RNAPIIS5p-S7p regions. It would be of interest to understand whether regions more transcribed have some other transcriptional feature compared to other regions, such as unidirectional versus bidirectional transcription.

I find that extragenic RNAPII enhancer regions are sensitive to Erk2 knock out (KO). Erk2 is known to regulate poised genes (Jia et al., 2012), while enhancers in different activation states are sensitive to Erk2 KO. Erk2 occupancy at extragenic regions is very low, however the effect on total RNAPII depletion after Erk2 KO suggests either an indirect effect and that RNAPII binding at enhancer regions might be less stable than at gene, therefore perturbation of its regulatory machinery could have stronger effect. For example, Cdk8 and Med12, two proteins involved in transcription re-initiation, are not enriched at extragenic enhancers (data not shown), which could be one of the reasons of less stable RNAPII binding and transcription at enhancers. More studies on the players involved in RNAPII recruitment at enhancer regions and transcription regulators would clarify whether transcription at promoters and enhancers is initiated and regulated similarly and also help understand the nature of transcriptionally active enhancers.

Enhancer and promoters share substantial similarities

Recently, active promoters were shown to transcribe a short-lived species of RNA on the reverse strand (Duttke et al., 2015; Flynn et al., 2011). It is still debated if all active promoters show bi-directional transcription (Duttke et al., 2015; Lepoivre et al., 2013), however it was shown that the two transcription events originate from distinct regions, 180bp apart (Andersson et al., 2014b) which suggests that reverse transcription is not due to wrong orientation of RNAPII at the promoter; remarkably upstream-transcribed regions are also occupied by RNAPIIS2p (Preker et al., 2011). Transcripts originating from upstream regions are short, not-polyadenilated and Exosome sensitive (Andersson et al., 2014b; Flynn et al., 2011; Pefanis et al., 2014), remarkably, these features are shared with eRNAs. Bidirectional transcription was described as a feature of

enhancer regions (Andersson et al., 2014a) and enhancers show similar pre-initiation complex (PIC) formation to promoters (Core et al., 2014; Scruggs et al., 2015).

I show that RNAPII states at enhancers are similar to the one found at promoters, however in most cases enhancer transcripts do not reach full maturation. Possible differences in the RNA maturation could be derived by the differential recruitment of RNA maturation machinery, which are still to be investigated. The current work and others (Hah et al., 2013; Henriques et al., 2018) have shown that players active at coding regions, such as Spt5, Erk2, and Cdk9 are also involved at enhancer to regulate transcription.

Recent work on promoter upstream regions showed premature termination due to non-sense or early termination sites (Ntini et al., 2013). Therefore, transcription at upstream regions can be regulated similarly to active coding region, however the transcription will be aborted after the encounter with a non-sense termination site. This finding, together with the finding that bidirectional promoters can act as enhancers (Dao et al., 2017; Mikhaylichenko et al., 2018) and enhancer can act as weak promoters (Arnold et al., 2013; Mikhaylichenko et al., 2018; Serfling et al., 1985) suggest a theory where promoters are specialised enhancers, which are followed downstream by an evolutionary conserved region, that is the coding gene (similar model proposed in Andersson *et al.* (Andersson, 2014)).

In line with this theory, I find that RNAPII states and RNAPII regulation to be similar between enhancers and promoters, notably with comparable levels of nascent RNA. It was also shown that evolutionary younger promoters tend to have a bi-directional transcription that stabilise to unidirectional during co-evolution with the coding regions (Jin et al., 2017). One could speculate that with time, the transcribed region of the enhancer could potentially mutate and the eRNA acquire a function. This could then lead to an evolutionarily constrain of the enhancer regions, similarly to what happens with coding genes. It would be of interest to investigate whether more evolutionary conserved enhancers tend to have unidirectional transcription and whether eRNAs originating from these regions are more likely to have a function.

Polycomb extragenic regions are similar to Polycomb repressed genes

Enhancers marked by the Polycomb mark H3K27me3 attracted major interest for their role on the regulation of neuronal genes (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). In my analysis of extragenic regions, I also focused on Polycomb enhancer regions, their transcription, and their dynamics during neuronal differentiation. The neuronal differentiation datasets were an

optimal choice for the study of poised enhancers as they have been previously described in the neuronal differentiation pathway (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). In my study I show that Polycomb extragenic regions in mESC tend to acquire RNAPII in later stages of differentiation and become active, while their inactivation was a less likely event. Polycomb extragenic regions also showed a preferential localization in proximity to repressed genes. This finding correlates with previous studies, showing that poised enhancers tend to be closer to repressed genes (Koenecke et al., 2017; Rada-Iglesias et al., 2011). This preferential localization might be connected with poised enhancer function. For example, poised enhancers might be upstream of the target gene promoter, inside a repressive locus. Interestingly, it was shown that Polycomb repressed regions can reside in extended repressive loci (Ferrai et al., 2017; Koenecke et al., 2017). When the target gene needs to be activated, RNAPII could be recruited to the poised enhancer region, track the DNA to find the promoter and start productive transcription at coding genes.

Interestingly, extragenic Polycomb regions shared features described at Polycomb repressed genes. H3K27me3 can be found together with RNAPIIS5p and RNAPIIS7p, as was describe at genes (Brookes et al., 2012; Ferrai et al., 2017; Stock et al., 2007). Fascinatingly, I found that Polycomb repressed states at extragenic regions are also present in fully differentiated dopaminergic neurons, as was recently described for genic regions in Ferrai *et al.* 2017 (Ferrai et al., 2017). This observation is in partial contrast with the enhancer field, where poised enhancers are usually described as ESC specific (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011), however low numbers of poised enhancer were previously reported in other studies and not explored (Cruz-Molina et al., 2017). It would be of interest to investigate the target genes of poised enhancer in fully differentiated neurons, to understand for example whether they regulate recently identified Polycomb-repressed genes encoding for TFs not related with the neuronal lineage (Ferrai et al., 2017). Furthermore, It would be of great interest to understand if enhancer poising is specific for the neuronal lineage as currently hypothesized (Creyghton et al., 2010; Rada-Iglesias et al., 2011) or can be found in other differentiated cells.

8. Bibliography

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature reviews Genetics *13*, 720-731.

Akhtar, A., and Gasser, S.M. (2007). The nuclear envelope and transcriptional control. Nature reviews Genetics *8*, 507-517.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nature Reviews Molecular Cell Biology *16*, 155-166.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Developmental cell *16*, 47-57.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) *31*, 166-169.

Anderson, E., Devenney, P.S., Hill, R.E., and Lettice, L.A. (2014). Mapping the Shh long-range regulatory domain. Development (Cambridge, England) *141*, dev.108480--dev.108480-.

Anderson, E., and Hill, R.E. (2014). Long range regulation of the sonic hedgehog gene. Current opinion in genetics & development *27*, 54-59.

Andersson, R. (2014). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. BioEssays *37*, 314-23.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014a). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014b). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nature communications *5*, 5336-5336.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andrey, G., and Mundlos, S. (2017). The three-dimensional genome: regulating gene expression during pluripotency and development. Development *15*, 3646-3658.

Andrey, G., Schöpflin, R., Jerković, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M.*, et al.* (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome research *27*, 223-233.

Ang, S.L. (2006). Transcriptional control of midbrain dopaminergic neuron development. Development *133*, 3499-3506.

Aran, D., Abu-Remaileh, M., Levy, R., Meron, N., Toperoff, G., Edrei, Y., Bergman, Y., Hellman, A., Hanahan, D., Weinberg, R.A.*, et al.* (2016). Embryonic Stem Cell (ES)-Specific Enhancers Specify the Expression Potential of ES Genes in Cancer. PLoS genetics *12*, e1005840-e1005840.

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M.*, et al.* (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science *347*, 1010-1014.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science *339*, 1074-1077.

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

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299-308.

Baumli, S., Lolli, G., Lowe, E.D., Troiani, S., Rusconi, L., Bullock, A.N., Debreczeni, J.E., Knapp, S., and Johnson, L.N. (2008). The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. Embo j *27*, 1907-1918.

Beagrie, R.A., and Pombo, A. (2016). Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. BioEssays *38*, 881-893.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nature reviews Genetics *15*, 163-175.

Benton, M.L., Talipineni, S.C., Kostka, D., and Capra, J.A. (2017). Genome-wide Enhancer Maps Differ Significantly in Genomic Distribution, Evolution, and Function. bioRxiv *176610*

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., *et al.* (2006a). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.
Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., and Allis, C.D. (2006b). Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Molecular and cellular biology *26*, 2560-2569.
Blinka, S., Reimer, Michael H., Pulakanti, K., and Rao, S. (2016). Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes, Cell Reports *27,* 19-28
Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., *et al.* (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. (2009). Modifications of RNA polymerase II are pivotal in regulating gene expression states. EMBO reports *10*, 1213-1219.
Brookes, E., and Pombo, A. (2012). Code breaking: the RNAPII modification code in pluripotency. Cell cycle (Georgetown, Tex) *11*, 1267-1268.
Bu, H., Gan, Y., Wang, Y., Zhou, S., and Guan, J. (2017). A new method for enhancer prediction based on deep belief network. BMC Bioinformatics *18*, 418.
Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., and Wysocka, J. (2014). Reorganization of enhancer patterns in transition from naive to primed pluripotency. Cell stem cell *14*, 838-853.
Bulger, M., and Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. Genes Dev *13*, 2465-2477.
Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Molecular cell *36*, 541-546.
Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Molecular cell *49*, 825-837.
Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., *et al.* (2018). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res *46*, D762-d769.
Chapman, R.D., Heidemann, M., Hintermair, C., and Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. Trends in Genetics *24*, 289-296.
Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., *et al.* (2013). Discovery and Characterization of Super-Enhancer Associated Dependencies in Diffuse Large B-Cell Lymphoma. Cancer Cell *24*, 777-790.
Chen, C.-y., Morris, Q., and Mitchell, J.A. (2012). Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. BMC genomics *13*, 152-152.
Chen, F.X., Xie, P., Collings, C.K., Cao, K., Aoi, Y., Marshall, S.A., Rendleman, E.J., Ugarenko, M., Ozark, P.A., Zhang, A., *et al.* (2017). PAF1 regulation of promoter-proximal pause release via enhancer activation. Science *357*, 1294-1298.
Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., *et al.* (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.
Choi, I., Kim, R., Lim, H.-W., Kaestner, K.H., and Won, K.-J. (2014). 5-hydroxymethylcytosine represses the activity of enhancers in embryonic stem cells: a new epigenetic signature for gene regulation. BMC genomics *15*, 670-670.
Conaway, R.C., and Conaway, J.W. (2011). Function and regulation of the Mediator complex. Curr Opin Genet Dev *21*, 225-230.
Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.
Conway, J.R., Department of Biomedical Informatics, H.M.S., Boston, MA 02115, USA, Lex, A., SCI Institute, S.o.C., University of Utah, Salt Lake City, UT 84112, USA, Gehlenborg, N., and Department of Biomedical Informatics, H.M.S., Boston, MA 02115, USA (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938-2940.
Corden, J.L. (2013). RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chemical reviews *113*, 8423-8455.
Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics *46,* 1311-1320.
Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., *et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.
Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W.F.J., Grosveld, F., Frommolt, P., Bazzi, H., *et al.* (2017). PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. Cell Stem Cell *20*, 689-705.e689.

Danko, C.G., Hyland, S.L., Core, L.J., Martins, A.L., Waters, C.T., Lee, H.W., Cheung, V.G., Kraus, W.L., Lis, J.T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. Nature Methods *12*, 433-438.
Dao, L.T.M., Galindo-Albarrán, A.O., Castro-Mondragon, J.A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T.*, et al.* (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. Nature Genetics *49*, 1073-1081.
David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes & development *25*, 972-983.
de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome Res *11*, 447-459.
De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology *8*, e1000384-e1000384.
Descostes, N., Heidemann, M., Spinelli, L., Schüller, R., Maqbool, M.A., Fenouil, R., Koch, F., Innocenti, C., Gut, M., Gut, I.*, et al.* (2014). Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. eLife *3*, e02105-e02105.
Dias, J.D., Rito, T., Torlai Triglia, E., Kukalev, A., Ferrai, C., Chotalia, M., Brookes, E., Kimura, H., Pombo, A., Akhtar, M.S.*, et al.* (2015). Methylation of RNA polymerase II non-consensus Lysine residues marks early transcription in mammalian cells. eLife *4*, 387-393.
Dixon, J.R.Jesse R., Gorkin, David U.D.U., Ren, B., Alipour, E., Marko, J.F., Austenaa, L.M., Barozzi, I., Simonatto, M., Masella, S., Chiara, G.D.*, et al.* (2016). Chromatin Domains: The Unit of Chromosome Organization. Molecular Cell *62*, 668-680.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) *29*, 15-21.
Dowen JM, B.S., Orlando DA, Hübner MR, Abraham BJ, Spector DL , Young RA (2018). Multiple Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements. Stem Cell Reports *24,* 371-8.
Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R.*, et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.
Duttke, Sascha H.C., Lacadie, Scott A., Ibrahim, Mahmoud M., Glass, Christopher K., Corcoran, David L., Benner, C., Heinz, S., Kadonaga, James T., and Ohler, U. (2015). Human Promoters Are Intrinsically Directional. Molecular Cell *19,* 674-684.
Egloff, S., and Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends in genetics : TIG *24*, 280-288.
Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature.
Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods *9*, 215-216.
Estarás, C., Benner, C., and Jones, K.A. (2015). SMADs and YAP Compete to Control Elongation of β-Catenin:LEF-1-Recruited RNAPII during hESC Differentiation. Molecular cell *58,* 780-93.
Fabrega, C., Shen, V., Shuman, S., and Lima, C.D. (2003). Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol Cell *11*, 1549-1561.
Ferrai, C., and Pombo, A. (2009). 3D chromatin regulation of Sonic hedgehog in the limb buds. Developmental cell *16*, 9-11.
Ferrai, C., Triglia, E.T., Risner- Janiczek, J.R., Rito, T., Rackham, O.J., Santiago, I.d., Kukalev, A., Nicodemi, M., Akalin, A., Li, M.*, et al.* (2017). RNA polymerase II primes Polycomb- repressed developmental genes throughout terminal neuronal differentiation. Molecular Systems Biology *16,* 946.
Ferrari, K.J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stützer, A., Fischle, W., Bonaldi, T., and Pasini, D. (2014). Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. Molecular cell *53*, 49-62.
Fisher, R.P. (2017). CDK regulation of transcription by RNAP II: Not over 'til it's over? Transcription *15*, 81-90.
Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proceedings of the National Academy of Sciences of the United States of America *108*, 10460-10465.
Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S.*, et al.* (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Molecular systems biology *11*, 852-852.
Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.-L.*, et al.* (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity *32*, 317-328.
Ghosh, A., Shuman, S., and Lima, C.D. (2011). Structural insights to how mammalian capping enzyme reads the CTD code. Mol Cell *43*, 299-310.

Gosselin, D., Link, V.M., Romanoski, Casey E., Fonseca, Gregory J., Eichenfield, Dawn Z., Spann, Nathanael J., Stender, Joshua D., Chun, Hyun B., Garner, H., Geissmann, F.*, et al.* (2014). Environment Drives Selection and Function of Enhancers Controlling Tissue-Specific Macrophage Identities. Cell *159*, 1327-1340.
Grzechnik, P., Tan-Wong, S.M., and Proudfoot, N.J. (2014). Terminate and make a loop: regulation of transcriptional directionality. Trends Biochemical Science *39*, 319-327.
Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. Nucleic acids research *41*, 1591-1603.
Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome research *23*, 1210-1223.
Hahn, S., and Young, E.T. (2011). Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics *189*, 705-736.
Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2016). CTCF and Cohesin Regulate Chromatin Loop Stability with Distinct Dynamics. bioRxiv *093476*.
Harmanci, A., Rozowsky, J., and Gerstein, M. (2014). MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome Biology *15*, 474-474.
Hay, D., Hughes, J.R., Babbs, C., Davies, J.O., Graham, B.J., Hanssen, L., Kassouf, M.T., Marieke Oudelaar, A.M., Sharpe, J.A., Suciu, M.C.*, et al.* (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat Genet *48*, 895-903.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009a). Histone Modifications at Human Enhancers Reflect Global Cell Type-Specific Gene Expression. Nature *459*, 108-112.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009b). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.
Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144-154.
Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev *32,* 26-41.
Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E.*, et al.* (2012). Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. In EMBO J, pp. 2784-2797.
Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.
Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., Lee, T.I., Bradner, J.E., and Young, R.A. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Molecular cell *58*, 362-370.
Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A.*, et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, aad9024-aad9024.
Hsin, J.-P., Li, W., Hoque, M., Tian, B., and Manley, J.L. (2014). RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. eLife *3*, e02112-e02112.
Hsin, J.-P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes & development *26*, 2119-2137.
Hsin, J.-P., Sheth, A., and Manley, J.L. (2011). RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing. Science (New York, NY) *334*, 683-686.
Ing-Simmons E, Seitan VC, Faure AJ, Flicek P, Carroll T, Dekker J, Fisher AG, Lenhard B, Merkenschlager M. 2015. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res *25*: 504–513.Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.-S., Brady, J.N., Ozato, K., Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N.*, et al.* (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.
Jia, J., Zheng, X., Hu, G., Cui, K., Zhang, J., Zhang, A., Jiang, H., Lu, B., Yates, J., Liu, C.*, et al.* (2012). Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. Cell *151*, 576-589.
Jin, Y., Eser, U., Struhl, K., and Churchman, L.S. (2017). The Ground State and Evolution of Promoter Region Directionality. Cell *170*, 889-898.
Johnson, K.R., Gagnon, L.H., Tian, C., Longo-Guess, C.M., Low, B.E., Wiles, M.V., and Kiernan, A.E. (2018). Deletion of a Long-Range Dlx5 Enhancer Disrupts Inner Ear Development in Mice. Genetics *208*, 1165-1179.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife *3*, e02407-e02407.
Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.
Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F.*, et al.* (2008). The UCSC Genome Browser Database: 2008 update. Nucleic Acids Research *36,* 773-9.
Kashyap, V., Rezende, N.C., Scotland, K.B., Shaffer, S.M., Persson, J.L., Gudas, L.J., and Mongan, N.P. (2009). Regulation of Stem Cell Pluripotency and Differentiation Involves a Mutual Regulatory Circuit of the Nanog, OCT4, and SOX2 Pluripotency Transcription Factors With Polycomb Repressive Complexes and Stem Cell microRNAs. In Stem Cells Dev *18*,1093-1108.
Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S.*, et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Kim, Y.W., Lee, S., Yun, J., and Kim, A. (2015). Chromatin looping and eRNA transcription precede the transcriptional activation of gene in the β-globin locus. Bioscience reports *18,* 35.
King, H.W., and Klose, R.J. (2017). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife *6*, e22631.
Kleftogiannis, D., Kalnis, P., and Bajic, V.B. (2014). DEEP: a general computational framework for predicting enhancers. Nucleic acids research *43*, e6--e6-.
Knuesel, M.T., Meyer, K.D., Bernecky, C., and Taatjes, D.J. (2009). The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. Genes Dev *23*, 439-451.
Ko, J.Y., Oh, S., and Yoo, K.H. (2017). Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. Mol Cells *40*, 169-177.
Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C.*, et al.* (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology *18*, 956-963.
Koenecke, N., Johnston, J., He, Q., Meier, S., and Zeitlinger, J. (2017). Drosophila poised enhancers are generated during tissue patterning with the help of repression. Genome research *27*, 64-74.
Krivega, I., and Dean, A. (2012). Enhancer and promoter interactions — long distance calls. Curr Opin Genet Dev *22*, 79-85.
Kvon, E.Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. Genomics *106*, 185-192.
Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497-501.
Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M.*, et al.* (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511-515.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.
Lawrence , M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2018). Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology.
Le Gras, S., Keime, C., Anthony, A., Lotz, C., De Longprez, L., Brouillet, E., Cassel, J.-C., Boutillier, A.-L., and Merienne, K. (2017). Altered enhancer transcription underlies Huntington's disease striatal transcriptional signature. Scientific Reports *7*, 42875-42875.
Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.-A., Koch, F., Maqbool, M.*, et al.* (2013). Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics *14*, 914-914.
Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2018). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics *12*, 1725-1735.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754-1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078-2079.
Li, Q., Peterson, K.R., Fang, X., and Stamatoyannopoulos, G. (2002). Locus control regions. Blood *100*, 3077-3086.
Li, W., Hu, Y., Oh, S., Ma, Q., Merkurjev, D., Song, X., Zhou, X., Liu, Z., Tanasa, B., He, X.*, et al.* (2015). Condensin I and II Complexes License Full Estrogen Receptor α-Dependent Enhancer Activation. Molecular cell *59*, 188-202.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., *et al.* (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516-520.
Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nature Reviews Genetics *17* , 207-23.
Liber D, Domaschenz R, Holmqvist PH, Mazzarella L, Georgiou A, Leleu M, Fisher AG, Labosky PA, Dillon N. (2010). Epigenetic Priming of a Pre-B Cell-Specific Enhancer through Binding of Sox2 and Foxd3 at the ESC Stage. Cell Stem Cell *7*, 114-126.
Liu, W., Ma, Q., Wong, K., Li, W., Ohgi, K., Zhang, J., Aggarwal, A.K., and Rosenfeld, M.G. (2013). Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell *155*, 1581-1595.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550-550.
Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.
Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nat Struct Mol Biol *17*, 1195-1201.
Lupiáñez, Darío G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, John M., Laxova, R., *et al.* (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell *161*, 1012-1025.
Ma, Y., Yuan, J., Huang, M., Jove, R., and Cress, W.D. (2003). Regulation of the Cyclin D3 Promoter by E2F1.
Markenscoff-Papadimitriou, E., Allen, William E., Colquitt, Bradley M., Goh, T., Murphy, Karl K., Monahan, K., Mosley, Colleen P., Ahituv, N., and Lomvardas, S. (2014). Enhancer Interaction Networks as a Means for Singular Olfactory Receptor Expression. Cell *159*, 543-557.
Matarese, F., Carrillo-de Santa Pau, E., and Stunnenberg, H.G. (2011). 5-Hydroxymethylcytosine: a new kid on the epigenetic block? In Mol Syst Biol, p. 562.
Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012a). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science (New York, NY) *336*, 1723-1725.
Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012b). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science *336*, 1723-1725.
McKinney, W. (2018). Data Structures for Statistical Computing in Python. Paper presented at: Proceedings of the 9th Python in Science Conference.
McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.
Melo, C.A., Drost, J., Wijchers, P.J., van de Werken, H., de Wit, E., Oude Vrielink, J.A., Elkon, R., Melo, S.A., Leveille, N., Kalluri, R., *et al.* (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell *49*, 524-535.
Merika, M., and Thanos, D. (2001). Enhanceosomes. Curr Opin Genet Dev *11*, 205-208.
Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription.
Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49-55.
Milewski, R.C., Chi, N.C., Li, J., Brown, C., Lu, M.M., and Epstein, J.A. (2004). Identification of minimal enhancer elements sufficient for Pax3 expression in neural crest and implication of Tead2 as a regulator of Pax3. Development *131*, 829-837.
Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, Glass CK (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular cell *51*, 310-325.
Morris, K.J. (2012). Interplay between Polycomb repression and RNA Polymerase II in embryonic stem cells. PhD Thesis. Imperial College London.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods *5*, 621-8.
Nord, A.S., Blow, M.J., Attanasio, C., Akiyama, J.A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., *et al.* (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell *155*, 1521-1531.
Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745.
Oishi, Y., Hayashi, S., Isagawa, T., Oshima, M., Iwama, A., Shimba, S., Okamura, H., and Manabe, I. (2017). Bmal1 regulates inflammatory responses in macrophages by modulating enhancer RNA transcription. Scientific Reports *7*, 7086.
Orom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., *et al.* (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell *143*, 46-58.
Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., *et al.* (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239.
Palstra, R.J., and Grosveld, F. (2012). Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. Front Genet *3,* 195.
Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A., Rico, D., Sanborn, A.L., Rao, S.S.P., *et al.* (2016). Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biology *17*, 152-152.
Paralkar, Vikram R., Taborda, Cristian C., Huang, P., Yao, Y., Kossenkov, Andrew V., Prasad, R., Luan, J., Davies, James O.J., Hughes, Jim R., Hardison, Ross C., *et al.* (2016). Unlinking an lncRNA from Its Associated cis Element. Molecular Cell *62*, 104-110.
Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., Bueren, K.L.v., Chines, P.S., Narisu, N., Program, N.C.S., *et al.* (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci USA *110,* 17921-17926.
Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature *514*, 389-393.
Plank, J.L., and Dean, A. (2014). Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. Molecular Cell *55*, 5-14.
Pradeepa, M.M. (2016). Causal role of histone acetylations in enhancer function. Transcription *8,* 40-47.
Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C.A., Schneider, R., and Bickmore, W.A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. Nature genetics *48*, 681-686.
Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res *39*, 7179-7193.
Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science (New York, NY) *322*, 1851-1854.
Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. Science *352*, aad9926.
Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841-842.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.
Rahl, P.B., Whitehead Institute for Biomedical Research, C., MA 02142, USA, Lin, C.Y., Whitehead Institute for Biomedical Research, C., MA 02142, USA, Department of Biology, M.I.o.T., Cambridge, MA 02142, USA, Seila, A.C., Koch Institute, M.I.o.T., Cambridge, MA 02142, USA, Flynn, R.A., Koch Institute, M.I.o.T., Cambridge, MA 02142, USA, McCuine, S., *et al.* (2010). c-Myc Regulates Transcriptional Pause Release. Cell *141*, 432-445.
Rahman, S., Zorca, C.E., Traboulsi, T., Noutahi, E., Krause, M.R., Mader, S., and Zenklusen, D. (2016). Single-cell profiling reveals that eRNA accumulation at enhancer-promoter loops is not required to sustain transcription. Nucleic acids research. Nucleic Acids Res *45,* 3017-3030.
Ramírez, F., Max Planck Institute of Immunobiology and Epigenetics, F., Germany, Ryan, D.P., Max Planck Institute of Immunobiology and Epigenetics, F., Germany, Grüning, B., University of Freiburg, D.o.C.S., 79110 Freiburg, Germany, Bhardwaj, V., Max Planck Institute of Immunobiology and Epigenetics, F., Germany, Faculty of Biology, U.o.F., 79104 Freiburg, Germany, Kilpert, F., *et al.* (2018). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research *44*.
Rao, Suhas S.P., Huntley, Miriam H., Durand, Neva C., Stamenova, Elena K., Bochkov, Ivan D., Robinson, James T., Sanborn, Adrian L., Machol, I., Omer, Arina D., Lander, Eric S., *et al.* (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665-1680.
Reeder, C., Closser, M., Poh, H.M., Sandhu, K., Wichterle, H., and Gifford, D. (2015). High Resolution Mapping of Enhancer-Promoter Interactions. PLoS One *10,* e0122420.

Beagrie RA, Scialdone A, Schueler M, Kraemer DC, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas LM, Branco MR, Fraser J, Dostie J, Game L, Dillon N, Edwards PA, Nicodemi M, Pombo A. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543,* 519-524.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology *29*, 24.
Roehr, J.T., Institute of Bioinformatics, D.o.M.a.C.S., FU Berlin, 14195 Berlin, Germany, Dieterich, C., Klaus Tschira (2018). Flexbar 3.0 – SIMD and multicore parallelization. Bioinformatics *33*, 2941-2942.
Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.
Schaukowitch, K., Joo, J.-Y., Liu, X., Watts, Jonathan K., Martinez, C., and Kim, T.-K. (2014a). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Molecular Cell *56*, 29-42.
Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.K. (2014b). Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell *56*, 29-42.
Schröder, S., Herker, E., Itzen, F., He, D., Thomas, S., Gilchrist, D.A., Kaehlcke, K., Cho, S., Pollard, K.S., Capra, J.A*., et al.* (2013). Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells. Molecular cell *52*, 314-324.
Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Molecular cell *58*, 1101-1112.
Serfling, E., Jasin, M., and Schaffner, W. (1985). Enhancers and eukaryotic gene transcription. Trends in Genetics *1*, 224-230.
Shin, H.Y., Willi, M., HyunYoo, K., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat Genet *48*, 904-911.
Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences *39*, 381-399.
Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. Embo j *11*, 4047-4057.
Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics *13*, 613.
Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform *17*, 953-966.
Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G., and Pombo, A. (2007). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nature cell biology *9*, 1428-1435.
Tee, W.-W., Shen, Steven S., Oksuz, O., Narendra, V., and Reinberg, D. (2014). Erk1/2 Activity Promotes Chromatin Features and RNAPII Phosphorylation at Developmental Promoters in Mouse ESCs. Cell *156*, 678-690.
Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc Natl Acad Sci U S A *110*, 18602-18607.
Thibodeau, A., Márquez, E.J., Shin, D.-G., Vera-Licona, P., and Ucar, D. (2017). Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. Scientific Reports *7*, 14466.
Thomas, M.C., and Chiang, C.M. (2006). The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol *41*, 105-178.
Thomas, R., Thomas, S., Holloway, A.K., and Pollard, K.S. (2017). Features that define the best ChIP-seq peak calling algorithms. Brief Bioinform *18*, 441-450.
Tietjen, J.R., Zhang, D.W., Rodríguez-Molina, J.B., White, B.E., Akhtar, M.S., Heidemann, M., Li, X., Chapman, R.D., Shokat, K., Keles, S*., et al.* (2010). Chemical-genomic dissection of the CTD code. Nature structural & molecular biology *17*, 1154-1161.
Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2012a). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology *31*, 46-53.
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012b). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.
Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, Edison T., Bühler, M*., et al.* (2013). Regulation of Transcription through Acetylation of H3K122 on the Lateral Surface of the Histone Octamer. Cell *152*, 859-872.
Van der Ploeg LH , e.a. (1980). gamma-beta-Thalassaemia studies showing that deletion of the gamma- and delta-genes influences beta-globin gene expression in man. - PubMed - NCBI. Nature.
Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T., Fernandez, N., Ballester, B., Andrau, J.C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun *6*, 6905.

Vernimmen, D., and Bickmore, W.A. (2015). The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. Trends Genet *31*, 696-708.
Vincent, T., Neve, E.P.A., Johnson, J.R., Kukalev, A., Rojo, F., Albanell, J., Pietras, K., Virtanen, I., Philipson, L., Leopold, P.L.*, et al.* (2009). A SNAIL1-SMAD3/4 transcriptional repressor complex promotes TGF-β mediated epithelial-mesenchymal transition. Nat Cell Biol *11*, 943-950.
Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res *35*, D88-92.
Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes & development *27*, 1318-1338.
Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci *131*, 281-285.
Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A.*, et al.* (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature *474*, 390-394.
Wang, Q., Carroll, J.S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol Cell *19*, 631-642.
Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.
Wickham, H. (2016). ggplot2 - Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4.
Wilkinson, L. (2012). Exact and approximate area-proportional circular Venn and Euler diagrams. IEEE Trans Vis Comput Graph *18*, 321-331.
Winter, G.E., Mayer, A., Buckley, D.L., Erb, M.A., Roderick, J.E., Vittori, S., Reyes, J.M., di Iulio, J., Souza, A., Ott, C.J.*, et al.* (2017). BET Bromodomain Proteins Function as Master Transcription Elongation Factors Independent of CDK9 Recruitment. Molecular Cell *67*, 5-18.e19.
Wotton, D., Lo, R.S., Lee, S., and Massague, J. (1999). A Smad transcriptional corepressor. Cell *97*, 29-39.
Xing, H., Mo, Y., Liao, W., and Zhang, M.Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS computational biology *8*, e1002613-e1002613.
Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes & development *23*, 2824-2838.
Yang, C., and Stiller, J.W. (2014). Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A *111*, 5920-5925.
Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2014). Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556-556.
Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev *25,* 2227-41.
Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome research *21*, 1273-1283.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137-R137.

9. Appendix

9.1 Permission for figures

Figure 1.2 is reproduced from "The three-dimensional genome: regulating gene expression during pluripotency and development", Development, 2017.

Requestor type	Author of requested content
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
The requesting person/organization	Giulia Caglio
Title or numeric reference of the portion(s)	Figure 1
Title of the article or chapter the portion is from	The three-dimensional genome: regulating gene expression during pluripotency and development
Author of portion(s)	Guillaume Andrey, Stefan Mundlos
Volume of serial or monograph	144
Page range of portion	3646-3658
Publication date of portion	October 17, 2017
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	RNA Polymerase II identifies enhancers in different states of activation
Expected presentation date	Apr 2018

Figure 1.4 is reproduced from "Modifications of RNA polymerase II are pivotal in regulating gene expression states", EMBO Reports, 2009.

This Agreement between MDC ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4332050912873
License date	Apr 18, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	EMBO Reports
Licensed Content Title	Modifications of RNA polymerase II are pivotal in regulating gene expression states
Licensed Content Author	Emily Brookes, Ana Pombo
Licensed Content Date	Oct 16, 2009

Licensed Content Volume	10
Licensed Content Issue	11
Licensed Content Pages	7
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 2
Will you be translating?	No
Title of your thesis / dissertation	RNA Polymerase II identifies enhancers in different states of activation
Expected completion date	Apr 2018
Expected size (number of pages)	250
Publisher Tax ID	EU826007151
Total	0.00 EUR

Figure 6.1 is reproduced from "RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation", Molecular System biology, 2017. This article is under a Creative Commons Licence (Attribution 3.0 Unported (CC BY 3.0)).