Poising and connectivity of enhancers upon naïve-to-primed transition in human embryonic stem cells (hESCs)

Monica Della Rosa

Imperial College London

Institute of Clinical Sciences

A thesis presented for the degree of Doctor of Philosophy and

Diploma of Imperial College London

Declaration of Authenticity and Copyright

I, Monica Della Rosa, declare that the work presented in this thesis is my own and did not use any unnamed sources. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made by correct citation. The work contained in this thesis has not been previously submitted for examination or any other degree.

Acknowledgments

Firstly, I would like to thank you Mikhail, for giving me the opportunity to work on this PhD project. Your enthusiasm, your genuine curiosity and your ability to always find the positive side of any setback have been invaluable. I've learned so much (besides R)!

I would like to thank the FGC lab. Especially, Lera and Helen! I couldn't have wished for better companions! I've been so lucky to being able to share this with you guys and make some great friends in the process.

A huge thank you goes to some very amazing people, for their help and their support throughout and for being part of a great support system: Eleni, Antoine, Radina, Sijia and all the people in ICTEM, thank you!

The amazing Imperialists! I'm so happy to have shared this experience with such a great bunch!

Jacopo, tu hai avuto un ruolo determinante - andando indietro nel tempo abbastanza, forse non sarei nemmeno arrivata qui se non ti avessi incrociato ad un certo punto della vita. Sei una persona speciale per me.

Grazia, come spero tu sappia, sei stata fondamentale in questo percorso - e lo sei tutto'ora! La tua tenacia, forza e determinazione sono per me un esempio e custodisco la nostra amicizia preziosamente.

My family. Mom and Dad, all my brothers and sisters and Daniela, Franci and Fabio. You gifted me with the freedom to choose who I wanted to be. Never judging, never questioning, always with support and always believing in me. Ed ovviamente, Andrea & Elia. I due puffetti che riempiono le stanze più grandi con la loro energia e dolcezza. Siete la cosa più preziosa!*"Always and forever"*.

A big thank you goes to Terry and Christine. Your constant support means so much to me. You make me feel at home, always.

Alex. Well, it's really hard to know where to begin. You caught every tear, shared every smile and I couldn't have done this without your support. You believed in me when I couldn't and there's no words to thank you for it!

Nadia, sei andata via un po' troppo presto...è un onore poterti chiamare "famiglia".

Finally, every single one of you who's been part of this journey!

Abstract

Enhancers are non-coding DNA elements that play crucial roles in transcriptional control, particularly in development. Patterns of histone modifications at enhancers are commonly used to infer their activity states and, poised enhancers (PEs) in particular display a 'bivalent' chromatin state: the 'active' H3K4me1 and the 'repressive', Polycomb-associated H3K27me3. Typically observed in pluripotent stem cells (PSCs), it was shown that PEs are required for gene activation later during differentiation. However, the function of the poised state of enhancers remains largely unknown.

To trace the emergence of PEs in early development, I have extensively optimized a recently developed low-cell number Capture Hi-C protocol to perform Poised Enhancer Capture Hi-C (PECHi-C) in PSCs, in time course upon the naïve-to-primed transition, which is known to associate with a major shift in the localisation of Polycomb proteins, from a broader to a highly focal pattern. PECHi-C revealed that the PE-mediated regulatory circuitry undergoes significant reorganization between the two states. In particular, I detected three predominant patterns of PE-mediated interactions: the **UP**, **DOWN** and **CONSTANT** interaction classes. Integrating these results with Cut&Tag data on histone modifications revealed an interplay between the acquisition of the poised state at enhancers and their interaction dynamics whereby, at least in some cases, the acquisition of the bivalent signature occurs in parallel to the acquisition of their contacts. Moreover, the analyses suggested that **day 3** of the transition is a pivotal point of the naïve-to-primed transition for the emergence of PEs.

Overall, this thesis provided further insights into the emergence of PE-mediated regulatory circuitry during early embryogenesis. The different patterns of PE connectivity suggest the presence of diverse regulatory mechanisms of PEs, further suggesting that PEs might play a role at earlier stages of embryogenesis, by ensuring the correct transition from the ground state of pluripotency to the primed state.

Contents

List of Figures

List of Tables

List of abbreviations

3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
5C	3C carbon copy
AEBP2	Adipocyte Enhancer-Binding Protein 2
ART	Assisted Reproductive Technology
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
BDs	Bivalent Domains
bp	base pair
BRD	Bromodomain containing protein
BSA	Bovin Serum albumin
CBP	CREB-binding protein
CHi-C	Capture Hi-C
ChIP-seq	Chromatin immuno-precipitation sequencing
chromEMT	Electron microscopy tomography
ConA	Concanavalin A
cPRC1	Canonical Polycomb Repressive Complex 1
CREs	Cis-regulatory elements
CTCF	CCCTC-binding factor
CTD	C-terminal domain
Cut&Run	Cleavage Under Targets and Release Using Nuclease
Cut&Tag	Cleavage Under Targets and Tagmentation
DE	Definitive endoderm
DHFR	Dehydrofolate reductase
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNase-seq	DNase-I hypersensitivity sites sequencing

DSIF	5,6-dichloro-1-b- dribofuranosylbenzimidazole sensitivity -inducing factor
EpiSCs	Epiblast-like stem cells
EPOP	Elonging B/C and PRC2-associated protein
ESCs	Embryonic Stem Cells
EZH	Enhancer of zeste
FISH	Fluorescence in-situ hybridization
GFP	Green fluorescent protein
GWAS	Genome-Wide Association Studies
HAT	Histone acetyltranferase
HiFi	High Fidelity
hNECs	human neuroectoderm cells
HSP60	Heat shock protein 60
ICM	Inner Cell Mass
iCRISPRa	inducible CRISPR-activation
IDRs	Intrinsically disordered regions
IP	Immuno-Precipitation
iPSCs	Induced pluripotent stem cells
IVF	*In Vitro* Fertilization
JARID2	Jumonji, AT Rich Interactive Domain 2
Kb	kilobase pair
LADs	Lamina-Associated Domains
LB	Luria Broth
LCR	Locus control region
LIF	Leukemia inhibitory factor
LLPS	Liquid-liquid phase separation
lncRNAs	Long non-conding RNAs
Mb	Megabase pair
Micro-C	Micro-Capture-C
MNase	Micrococcal nuclease
mRNA	Messanger RNA
ncPRC1	non-canonical Polycomb Repressive Complex 1
NDRs	Nucleosome depleted regions
NE	Neuroectoderm
NELF	Negative elongation factor

NGS	Next Generation Sequencing
nm	nanometer
nM	nano Molar
oCGIs	orphan CpG islands
PAL1/2	PRC2 Associated LCOR Isoform 1/2
PCA	Principal Component Analysis
PCGF	Polycomb group ring-finger domain protein
PcGs	Polycomb group proteins
PCHi-C	Promoter Capture Hi-C
PCLs	Polycomb-like homologs
PCR	Polymerase chain reaction
PECHi-C	Poised Enhancer Capture Hi-C
PEs	Poised Enhancers
PHC	Polyhomeotic homologous protein
PICs	Pre-initiation complexes
PRC1	Polycomb Repressive Complex 1
PRC2	Polycomb Repressive Complex 2
PREs	Polycomb Response Elements
PTMs	Post Translational Modifications
PWMs	Position Weight Matrices
RING1	Really Interesting New Gene 1
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNApolII	RNA polymerase II
RPKM	Reads per million mapped reads
rpm	rotation per minute
S2E	Even-skipped stripe 2 enhancer
SAM	Sterile alpha motif
SV40	Simian virus 40
SWI/SNF	Switch Sucrose Non-Fermentable
TADs	Topological Associated Domains
TALENS	Transcription activator-like effector nucleases
TBP	TATA-box binding protein
TDE1	Illumina Tagment DNA enzyme

TF	Transcription factor
TMP	Trimethoprim
TrxG	Trithorax group
TSS	Transcription Start Site
ZFN	Zinc Fingers
ZGA	Zigotic Gene Activation

1 Introduction

1.1 CHROMATIN STRUCTURE OF THE GENOME

The term chromatin was first introduced in 1882 by Walther Flemming to define the granular matter he could observe in the nucleus of the cells [1]. Today, chromatin is generally defined as DNA that is associated with proteins and RNA molecules tightly packaged in the nucleus of eukaryotic cells.

1.1.1 Chromatin structure and function

Initial *in vitro* assays suggested the existence of two main types of chromatin fibers that could be observed in the nucleus: the 30nm the 10nm structures. The 10nm conformation, commonly referred to as "beads on a string", consists of 147bp of DNA wrapped around nucleosomes [2, 3]. Nucleosomes typically consist of a core, or nucleosome core particle (NCP), containing histone octamers formed by four highly conserved histone proteins: H2A, H2B, H3, H4, which are flanked by strings of DNA, known as linker DNA. Specifically, histone octamers consist of a dimer formed by H2A/H2B that interacts with a tetramer formed by H3/H4 [4, 5, 6]. The nucleosome structure is intrinsically very dynamic and can adopt many alternative conformations allowing dynamic folding and unfolding of the nucleosomal DNA (**Figure 1.1**). Indeed, Transverse Relaxation Optimized SpectroscopY (TROSY)-NMR (Nuclear Magnetic Resonance) based studies, a methodology that allows the study of large molecules and complexes providing insights of the structure, dynamics, reaction state and chemical environment of such molecules and complexes, allowed to probe the role of histone proteins and their assembly in determining the plasticity of the NCP and in mediating the supercoiling of the DNA in the nucleus [7]. The precise mechanism with which nucleosomes can coil into the 30nm chromatin fibers still remains an area of active studies, although two main models have been proposed. The solenoid model proposes that nucleosomes are arranged linearly, in a solenoid-type

helix, whereby adjacent nucleosomes are connected by a bent DNA linker. Alternatively, the zig-zag model suggests that nucleosomes are stacked by going back and forth and are connected by a straight linker DNA [8, 9, 10]. *In vitro* chromatin reconstruction experiments have provided evidence that nucleosomes are more likely to display a zig-zag disposition which appears to be mainly dependent on the presence of the histone linker H1, suggesting a possible role for H1 in the formation of the more compacted 30nm fiber [11].

However, the evidence that supports the existence of the 30nm chromatin fibers is controversial. Studies using electron microscopy approaches have shown that the 30nm fibers are not a common feature of eukaryotic chromatin [12, 13], with the exception of chicken erythrocytes [14] and starfish sperm [15]. Chromatin fibers measured by electron microscopy appear mostly irregular, with most of them with a diameter ranging between 10nm and 30nm. Over the past two decades, chromosome conformation capture (3C-) based assays (described in more details in **section 1.1.5**) failed to provide precise information on the presence of 30nm fibers *in vivo* [16, 17, 18, 19]. However, recent studies based of the newly developed Micro-C (described in **section 1.1.5**) have mapped local chromatin folding at nucleosome resolution, providing evidence of the 30nm in yeast as well as in mammalian cells [20, 21].

Recently, the development of chromEMT (electron microscopy tomography) allowed visualisation of the chromatin ultra-structure and the 3D organization of the nucleus at multiple scales. Studies employing chromEMT found that only a minority of the chromatin is actually found in the 30nm structure ($< 20\%$), with most of it being in the "beads on a string" conformation. This suggests that the 10nm is the preferred conformation of chromatin under more physiological conditions [22].

Nevertheless, it is clear that one of the main roles of chromatin is to allow the necessary degree of compaction needed to store the DNA in the relatively small volume of the nucleus. However, DNA also needs to remain accessible for processes like DNA replication, DNA repair and transcription. Therefore, chromatin structure has to remain sufficiently dynamic to ensure prompt accessibility of the underlying DNA molecule.

Nucleosomes act as highly dynamic units that can be re-positioned along the DNA molecule allowing chromatin to change its shape and degree of compaction to rapidly move between different conformations with different degree of accessibility. Conventionally, we can distinguish two main types of chromatin: euchromatin and heterochro-

matin. Euchromatin usually refers to an open conformation, typically associated with actively transcribed regions. Heterochromatin, on the contrary, represents a more condensed state typically associated with gene-poor regions and low transcriptional activity [23]. We can further divide heterochromatin in constitutive and facultative. The former is mainly found over repetitive regions in the genome (such as telomeres, centromeres, etc.) and transposons while the latter typically encompasses genomic regions that need to be repressed in a cell-type specific manner [24, 25, 26, 27].

In general, the chromatin structure is highly complex and the development of novel technologies has allowed for greater characterisation of chromatin composition and spatial folding and how this ultimately impacts on gene regulation and other DNA-associated processes, which will be the focus of the following sections.

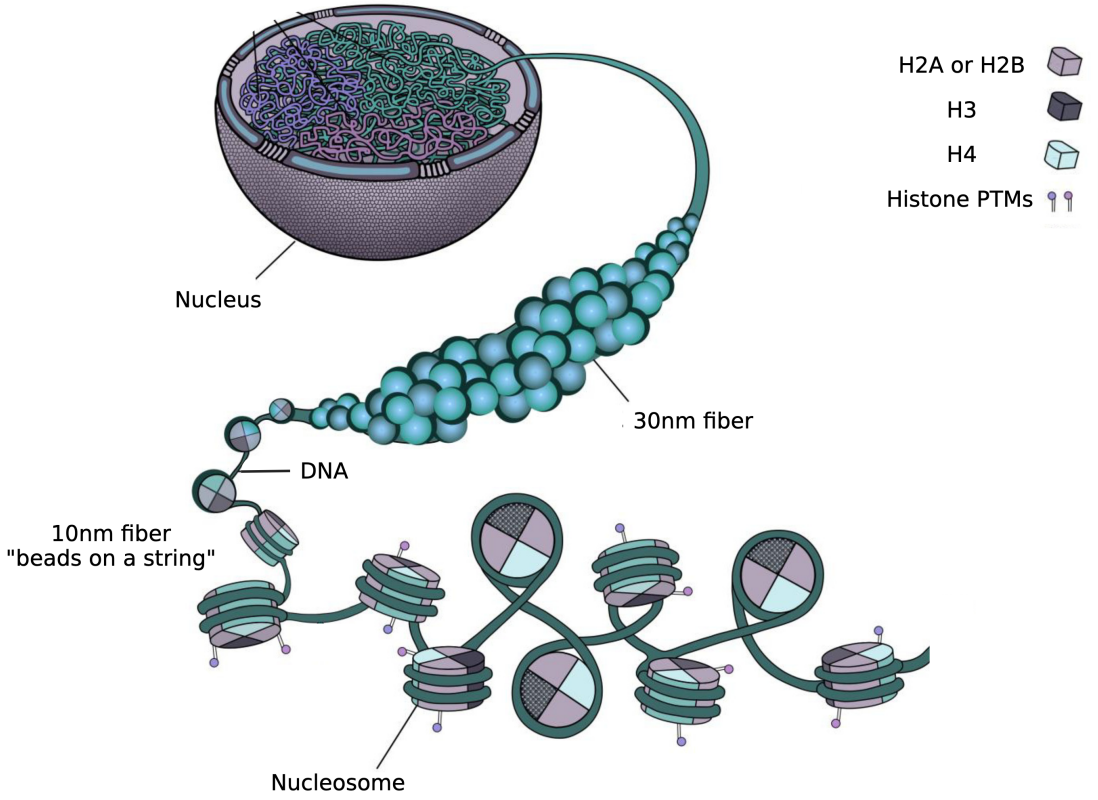


Figure 1.1: **Schematic of chromatin structure.** Schematic of the different levels of chromatin structure. Within the nucleus, DNA is wrapped around a histone octamer core, or nucleosome core particle (NCP), consisting of two copies of each histone: H2A, H2B, H3, H4. Two main types of chromatin can be observed: 10nm fiber, or "beads-on-a-string", and the 30nm fiber. Figure adapted from Rosa, S. & Shaw, P., 2013 [28].

1.1.2 Histone proteins and post-translational modifications (PTMs)

As previously mentioned, histone proteins form the NCP and they represent the most fundamental level of chromatin organization [29]. All histones share a similar structure, with a globular core domain and N-terminal "tails". The histone core domain forms the reel around which the DNA is wrapped and it represents approximately 75% of the histone protein mass. It is generally highly conserved from yeast to human and plays a key role in the maintenance of the genetic material. The remainder of the histone's mass is represented by their "tail" domain which, although structurally undefined, is highly conserved across evolution [30, 4]. Histone tails play a key role in the regulation of chromatin compaction. Indeed, *in vitro* studies show that their removal results in nucleosome arrays that are unable to condense into the 30nm fiber [31, 32, 33].

Histone N-terminal tails can undergo a series of different types of post-translational modifications (PTMs) that can mediate the regulation of different functional aspects of the chromatin [32]. In eukaryotes, one of the main role of PTMs is the regulation of DNA transcription [34, 35]. Typically, PTMs involve the addition of small groups such as the acetyl group (typically on lysine residues of histone tails), the methyl group (on lysine, arginine and glycine residues) and the phosphoryl group (typically found on serine and threonine), but, in some cases, they can also involve the addition of larger group such as ubiquitin or SUMO (Small Ubiquitin-like MOdifier) [30, 32, 36, 37, 38]. PTMs of histone tails, alone or in combination, can affect the interaction between histones and DNA and they can represent docking sites for numerous nuclear proteins [39].

In particular, histone acetylation and methylation of different lysines of histones H3 and H4 are among the most characterized PTMs and they play a crucial role in the regulation of transcriptional activation and repression.

Acetylation was initially associated with active gene transcription in the 1990s, with the purification of the histone acetyltranferase (HAT), HAT A, from *Tetrahymena* macronuclei [40, 32, 41]. HATs catalyze the transfer of an acetyl group to the ϵ-ammino group of histone tails lysine (K) residues and it is usually associated with chromatin accessibility and transcriptional activity (discussed in more details in **section 1.2**) [42, 43]. Histone deacetylases (HDACs) counteract HATs and are typically associated with decreased or absent levels of histone acetylation and transcriptional inactivity [44]. In general, histone acetylation has a central role in mediating a shift from a repressive to a permissive chro-

matin state through either promoting changes in the nucleosome structure or by creating docking sites for proteins involved in transcriptional regulation [43].

Methylation of histones involves the transfer of a methyl group to lysine, arginine or histidine residues of N-terminal tails. In particular, K can be methylated by one, two or three methyl groups (me1, me2, me3) [45, 46] and, depending on the specific K residue, methylation can be associated with transcriptional activity (e.g. H3K4, H3K36) or transcriptional repression and heterochromatin (e.g. H3K27, H3K9) [47, 48, 49].

Genome-wide studies to profile the main histone marks allowed to associate specific PTMs with a specific chromatin context [50]. These studies also identified a large cohort of domains where PTMs with seemingly opposite roles coexist and define regions characterized by either a bivalent or, is some cases, a trivalent signature. In both cases, it is thought that these more complex pattern of PTMs enable a rapid transition from an open to a closed chromatin state for a more precise regulation of transcription (bivalency will be described in more details in **section 1.3.4**) [51, 52].

In general, methylation of H3K4 and H3K27 are amongst the best characterized examples of methylation -associated transcriptional regulation [53]. Two major complexes have been identified to be responsible for the deposition of these marks. The Trithorax group (TrxG) is mainly responsible for the methylation of H3K4 and is more generally associated with transcriptional activation. Amongst the TrxG complexes, SWI/SNF (Switch Sucrose Non-Fermentable) acts mainly as chromatin remodeler and the COMPASS complex (a complex of protein associated with Set1) is the main responsible for the methylation of H3K4 [54]. On the opposite end, the Polycomb group proteins (PcGs) are responsible for the deposition of H3K27me3, typically associated with chromatin silencing and transcriptional repression. PcGs proteins form two main complexes: Polycomb Repressive complex 1 (PRC1) and PRC2 (described in more details in **section 1.1.3**) [55]. Both PcGs and TrxG are known to be highly conserved. Indeed, mutations in genes coding for both complexes are responsible for homeotic transformations in *Drosophila*, in support of their key role in coordinating proper gene expression patterns. Furthermore, both complexes have been well characterized for their role in the modification and modulation of chromatin [56, 57].

The next section will focus on PcGs and their role in mediating transcriptional repression.

1.1.3 Polycomb Repressive Complexes

Described in *Drosophila*, the *Polycomb* (*Pc*) gene was initially identified from the mutant phenotype associated with it, whereby heterozygous mutant male flies displayed additional sex comb teeth, hence the name Polycomb. It was then found that the effects given by the deletion of *Pc* were due to the aberrant regulation of homeotic genes [58, 59, 60]. In particular, it was observed that the *Pc* gene was responsible for the repression of the *Hox* cluster [58, 59]. Additional screening studies for suppressors of the *nos* phenotype, involved in the abdominal segmentation process in *Drosophila*, then highlighted the involvement of a specific component of the Polycomb group proteins (PcGs), enhancer of zeste 2 (EZH2), as a negative regulator of the abdomen-specific gap genes, *kni* and *gnt*. These studies showed that EZH2 was particularly involved in the maintenance of the anterior boundaries of both the *kni* and the *gnt* expression profiles, once their initial expression domains were set by the concentration gradient of the repressor Hunchback (Hb, the prime gap gene of the segmentation regulatory network in *Drosophila*) [60, 61]. It is now known that PcGs form highly conserved transcriptional repressive complexes known as PRCs. Two main complexes can be distinguished: PRC1 and PRC2 [62].

In mammals, **PRC1** complexes can be further subdivided into canonical PRC1 (cPRC1) and non-canonical PRC1 (ncPRC1). They all share a highly conserved core made by RING1 (Really Interesting New Gene) proteins (i.e. RING1A and RING1B), mainly responsible for the ubiquitylation of lysine 119 of histone H2A (H2AK119Ub) through their E3 ubiquitin ligase activity. RING1 proteins are usually coupled with one of the Polycomb group ring-finger domain proteins (PCGF1-PCGF6). Specifically, cPRC1 complexes are assembled around: PCGF2 and PCGF4, one of the chromobox proteins (CBX2, CBX4, CBX6-CBX8) that recognize and bind to H3K27me3 and/or H3K9me3 repressive marks and one of the Polyhomeotic (Ph) homologous proteins (PHC1-PHC3) characterized by a sterile alpha motif (SAM) domain that allows the formation of homo- or hetero-polymers of PcGs and is essential to achieve transcriptional repression [63]. ncPRC1 contains a zinc-finger domain, instead of CBX proteins, and a YY1-binding protein (RYBP or its paralog YAF2) which can associate with PCGF1/PCGF3, PCGF1/PCGF5 or PCGF6 to form, respectively, ncPRC1.1, ncPRC1.3, PRC1.5 or ncPRC1.6 [64] (**Figure 1.2**, top blue panel).

PRC2 mediates methyltransferase activity and the deposition of H3K27me3 [62]. Its main components are: SET-domain containing histone methylatransferases enhancer of zeste (EZH2 or EZH1), embryonic ectoderm development (EED), suppressor of zeste

(SUZ12) and the CAF1 histone binding proteins RBBP4 and RBBP7. Recent proteomic approaches in human revealed that the PRC2 core complex can itself be found in two main alternative assemblies: PRC2.1 and PRC2.2 [65]. In the PRC2.1 complex, the main core is bound to one of the three Polycomb-like homologs (PCLs), PHF1, PHF19 or MTF2, coupled with either EPOP (Elonging B/C and PRC2-associated protein) or PAL1/2 (PRC2 Associated LCOR Isoform 1/2) [66, 67]. Here, PHF1 (PHD finger protein 1) stimulates efficient trimethylation activity of EZH2 using as a substrate H3K27me2 that is recognized by its TUDOR domain [68]. It has also been shown that PHF1 and MTF2 mediate the recruitment of PRC2 at unmethylated CpGs [69, 70, 71, 72]. Moreover, the association of EPOP with several regions in mESCs led to speculate that it counteracts canonical PRC2 and transciptional silencing, maintaining sub-optimal transcription levels at bivalent genes (described in **section 1.3.4**) [73, 74, 75]. Meanwhile, PRC2.2 is defined by the presence of the zinc-finger proteins AEBP2 (Adipocyte Enhancer-Binding Protein 2) and JARID2 (Jumonji, AT Rich Interactive Domain 2) [65, 76, 77]. Both AEBP2 and JARID2 show preferential binding to CpGs (high CpG density regions in the genome) and they recognize the H2AK119Ub mark deposited by PRC1, potentially representing a functional link between the two main complexes [77, 78, 79, 80] (**Figure 1.2**, bottom red panel).

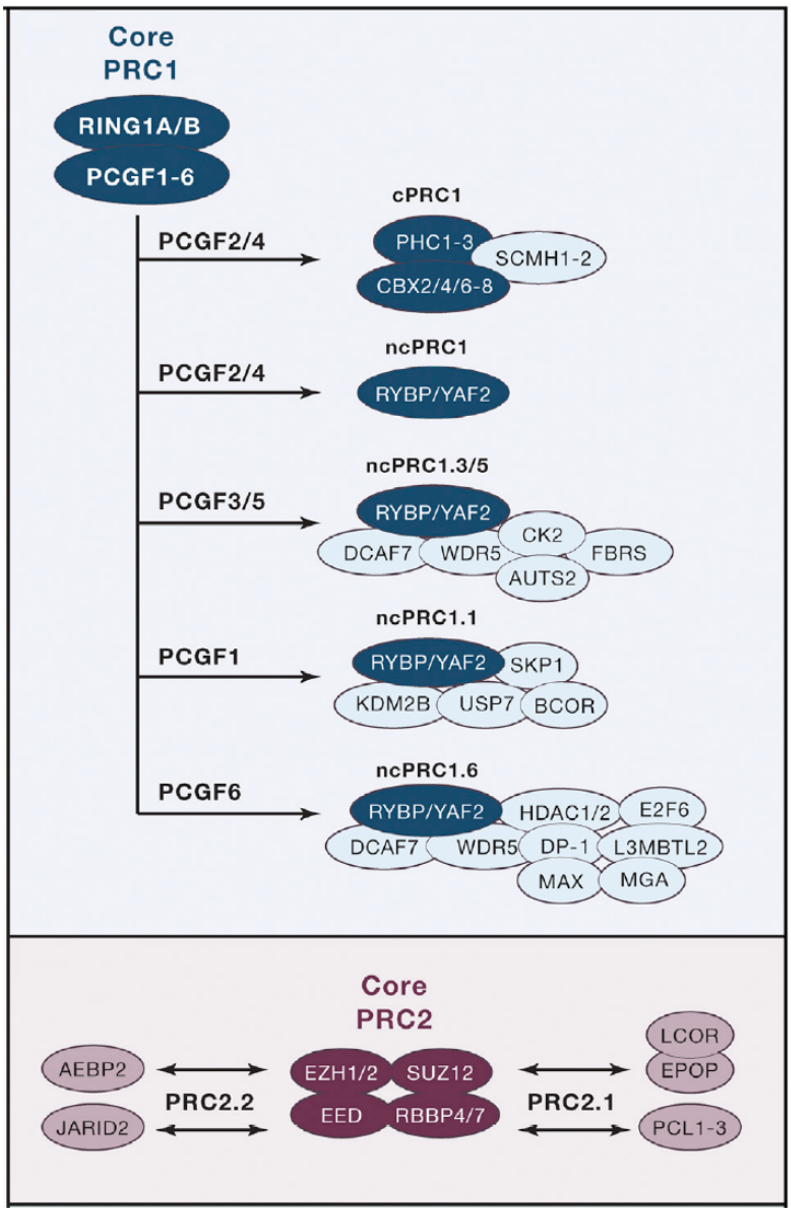


Figure 1.2: **Polycomb group protein complexes.** Polycomb group proteins (PcGs) constitute highly conserved transcriptional repressive complexes known as PRCs. Two main complexes can be distinguished: the PRC1 (blue panel) and the PRC2 complexes (red panel). The core PRC1 complex can associate with different PCGF proteins, PCGF1–PCGF6, forming the canonical PRC1 complexes (cPRC1) and non-canonical (ncPRC1) PRC1 complexes. The core of the PRC2 complex can associate with different accessory proteins to define the PRC2.1 and PRC2.2 complexes. Figure adapted from Schuettengruber, B., et al., 2017 [56].

Despite PcG proteins and their target genes being overall conserved during evolution [81], the number of PcG genes varies greatly across species and the study of the functional roles of the different PcGs complexes is still an area of active research. Nevertheless,

numerous studies have provided evidence that different PcG complexes are involved in different biochemical reactions and can target a variety of different genes, especially in embryonic stem cells (ESCs) [82, 83]. The requirements underlying the recruitment of PcGs can also diverge significantly. Early evidence in *Drosophila* showed that specific DNA regulatory elements, namely PREs (Polycomb Response Elements), can recruit PcGs factors and mediate the inheritance of silenced chromatin throughout development [84, 85, 86]. PREs were shown to contain transcription factor specific binding motifs and an initial model proposed that TFs recruited PRC2 in the first instance, which in turn recruited PRC1 by depositing H3K27me3 [87]. But while PREs play a crucial role in the recruitment of PcGs in *Drosophila*, they don't seem to be sufficient on their own. In fact, a more universal feature for PcGs recruitment in mammals is their preferential affinity to unmethylated CpG islands (CGIs). CGIs are usually defined as stretches of DNA, between 500bp-2000bp, with a CG content higher than 50% and an observed versus expected CpG ratio greater than 0.6. According to their methylation status, they can be associated with transcription repression (methylated CGIs) or activity (hypo- or un- methylated CGIs) [88, 89, 90].

In the context of PcGs recruitment, in particular, research has focused on KDM2B that binds to unmethylated CGIs through its CxxC-DNA binding domain playing an active role in the recruitment of PRC1.1 to the chromatin of ESCs. These findings led to the hypothesis that PRC1 tethers to chromatin first, leading to deposition of H3K119Ub. H3K119Ub then mediates the recruitment of PRC2 (perhaps through AEBP2 and JARID) which is responsible for the deposition of H3K27me3 [91, 92, 93].

Genome-wide mapping studies in mammals revealed that some of the PcG-bound CGIs correspond, indeed, to repressed promoters and absence of transcription can induce the recruitment of PcG proteins to these regions [92]. This supports the so called "chromatin sampling" model according to which PcG proteins weakly interact with all the potential binding sites present, but their binding is unstable if they encounter active transcription [94]. However, if transcription levels are low or absent, PcG proteins remain stably bound to CGIs. Although the "chromatin sampling" model seems plausible for most cases, PcG proteins can be also targeted to sites of active transcription, arguing for an alternative interrelated mechanism. Nevertheless, this implies that the cross-talk between PRC2 and PRC1 in the establishment of a repressive chromatin environment is more cooperative, rather than hierarchical, and highly context-dependent [56].

PcGs can mediate transcriptional repression through multiple mechanisms. One of the main and best characterized one is chromatin compaction. By inhibiting the chromatin-remodeling mediated by the SWI/SNF complex, PRC1 and PRC2 cooperatively render the chromatin inaccessible to the transcriptional machinery. It has recently been shown that a continuous competition between SWN/SNF and PcGs could be key in the switching between a repressed and an active chromatin state [95]. In addition to chromatin compaction, Polycomb can lead to transcriptional repression by inhibiting H3K27 acetylation. CBX proteins can, indeed, directly inhibit the HAT activity of CBP, therefore tipping the balance in favor of the repressive H3K27me3 [96].

Additionally, Polycomb activity can be regulated through its interaction with RNA [97, 98, 99]. Indeed, several studies have shown that PRC2 can promiscuously interact with different types of RNA: long non-coding RNAs (lncRNAs), short RNAs, transcribed by paused RNA polymerase II at the 5′ end of Polycomb-bound genes, or mRNAs of actively transcribed genes. The functional role of the Polycomb-RNA interaction has been an area of active research and several mechanisms have been explored. For example, it has been shown that RNA can inhibit the catalytic function of PRC2, particularly at regions devoid of H3K27me3, but less so at regions with pre-existing H3K27me3, playing a role in mediating the deposition and/or propagation of H3K27me3 [100, 101, 102, 103, 104, 105]. An alternative mechanisms of Polycomb-mediated silencing has been explored whereby the rixosome complex (involved in RNA processing and degradation) plays a role in the Polycomb-mediated gene repression through the interaction with PRC1. Specifically, RING1B recruits the rixosome complex at Polycomb-bound regions where RNA polymerase II is paused downstream of the promoter, and it promotes the degradation of nascent RNAs. This suggests that Polycomb-rixosome interaction might hinder productive transcriptional elongation by RNA polymerase II (**section 1.2.2**), ultimately leading to gene silencing [106, 107]. Overall, it has been suggested that RNA can have both a positive or a negative role in the regulation of Polycomb activity. It can either fine tune PRC2-mediated deposition of H3K27me3 or it can participate in mediating gene repression by the recrutiment of the rixosome RNA degradation complex through PRC1 and lead to impairment of transcriptional elongation [106].

High resolution profiles of both PRC1 and PRC2 and their respective histone marks H2AK119Ub and H3K27me3, particularly in ESCs, revealed that they can associate with large, inactive genomic loci, referred to as ”PcG bodies” or ”PcG clusters”. The first example of this was observed in *Drosophila* where the Bithorax complex (*BX-C*) gave evidence

that PREs physically interact with each other and with promoters [108, 109, 110]. Since then, multiple studies further confirmed the presence of PcG bodies, leading to the hypothesis that repression mediated by PcG occurs within nuclear substructures created by the folding of PcG-occupied chromatin [111].

1.1.4 3D spatial organization of the genome

Starting from a larger scale, 3D genome conformation studies highlighted the tendency of chromosomes within the nucleus to fold into two main compartments: a mostly active one, normally associated with a more open chromatin conformation, the **A compartment**, and a mainly inactive one which typically appears as more compacted chromatin, the **B compartment** [112]. A/B compartments are present on a scale of several megabases and, in general, regions within the same compartment tend to interact with one another [112] (**Figure 1.3**). While it is known that these compartments can be cell-type specific and can rearrange during development or in response to different gene expression patterns, they do not describe cell types in a comprehensive manner [18, 113].

While active genomic regions associate with one other at the center of the nucleus, inactive chromatin is usually located at the periphery of the nucleus and it is associated with the nuclear lamina (NL) through **LADs** (Lamina-Associated Domains, **Figure 1.3**). LADs are regions rich in heterochromatin and PTMs associated with gene repression. Similarly to heterochromatin, two types of LADs have been described: constitutive LADs and facultative LADs. The former are typically A-T rich and remain associated with the NL during differentiation, while the latter disengage when a gene becomes active [114, 115, 116, 117]. Although LADs associate with gene-poor regions or genes that are not being expressed, the mechanisms underlying the establishment of contacts between LADs and how this leads to gene repression still remain an area of active study [113].

On a 100kb to 800kb scale (and in some cases up to several megabases), genomic regions tend to interact with each other with higher frequency, within semi-confined clusters of chromatin interactions also known as Topologically Associated Domains or **TADs** (**Figure 1.3**). TADs represent a feature of the whole genome, they have been identified across different cell types and are conserved across species in the animal kingdom [118, 119, 120, 121, 113]. Although the mechanisms through which TADs are formed are not fully understood, it has been established that chromatin loop extrusion is involved in the formation of these structures. Loop extrusion is a process mediated predominantly by CCCTC-binding factor (CTCF). According to this model, chromatin loops are extruded

through cohesin rings until they encounter CTCF, which defines the boundaries of TADs [122, 123, 113] (**Figure 1.3**). The architectural role of TADs has been abundantly established, as well as their functional role in facilitating the regulation of precise transcriptional regulation (the role of TADs in the regulation of gene transcription will be explored further in **section 1.2.6**).

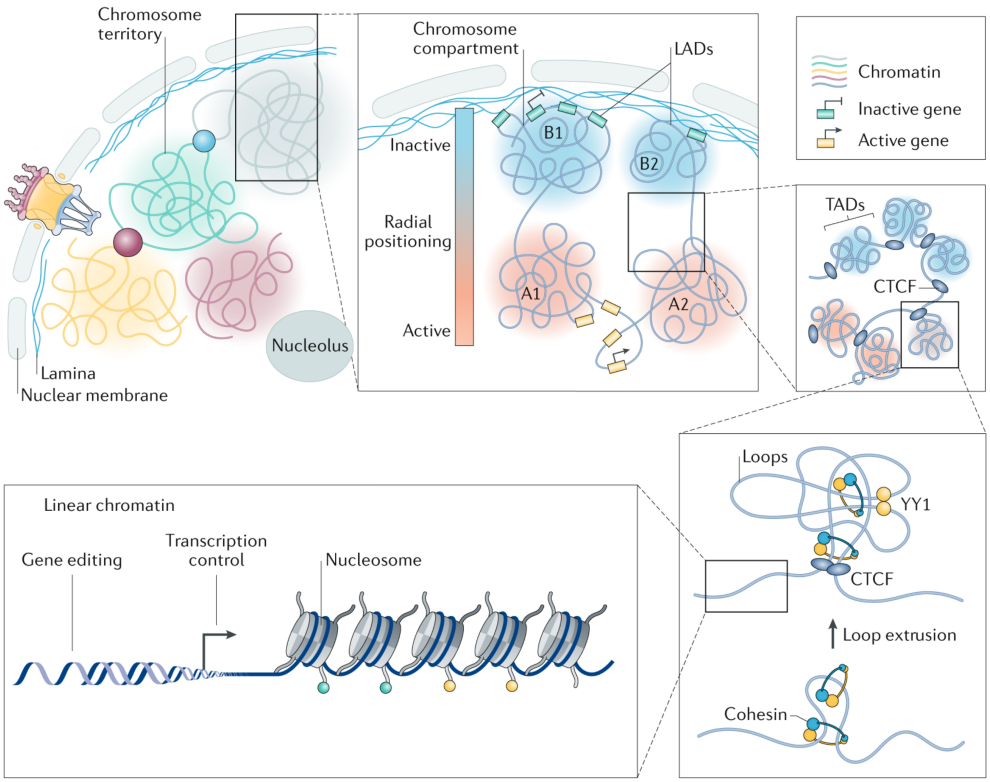


Figure 1.3: **Hierarchical organization of the 3D genome.** In the nucleus, chromosomes fold into two main compartments: the A compartment, usually located at the center of the nucleus and associated with active euchromatin. The B compartment, usually located at the periphery of the nucleus and associated with inactive heterochromatin. Lamina-associated domains (LADs)are transcriptionally inactive regions located close to the nuclear lamina. Chromatin regions that tend to interact with one another with higher frequency form topologically associating domains (TADs), separated by TAD boundaries enriched in structural proteins such as CTCF (CCCTC-binding factor) and cohesin. Based on the loop extrusion model, loop-extruding factors like cohesin actively extrude loops until they reach the boundary elements (i.e. CTCF-binding sites) and loop interactions can be stabilized by additional structural proteins such as CTCF, cohesin and YY1. The different levels of genome organization play a role in the fine regulation of gene expression. Figure adapted from Wang H. & Han M., et al., 2021 [124].

Over the last decade overwhelming evidence has shown that PcGs also play an im-

portant role in spatial folding of chromatin, forming repressive chromatin loops in the nucleus. The *Hox* cluster is a well known example of PcGs regulated regions displaying a looping structure [108, 125]. It has been hypothesized that H3K27me3 itself can stabilize PcGs-mediated 3D chromatin conformation contributing, therefore, to the fine regulation of their target genes [92, 111].

The role of PcGs in the 3D folding of the genome has been established in both flies and mammals. Especially in ESCs, several studies have led to hypothesize that CBX and PHC1/2 may contribute to the establishment of long-range interactions between PcG-bound loci and that PRC1 in cooperation with PRC2 may be involved [111]. 3D chromosome conformation studies in ESCs suggest that PcG bodies tend to localize in the A compartment where the chromatin is generally found in a more accessible state. A scenario that appears reversed in "terminally" differentiated cells, where PcG bodies shift toward the more inactive B compartment. This led to speculate that in ESCs Polycomb-bound chromatin places itself in a more permissive environment to be accessible to other TFs, co-activators and transcriptional machinery, therefore more responsive to activation cues [126, 127]. Moreover, it has recently been shown that the establishment of Polycomb-mediated long-range interactions is counteracted by cohesin through a mechanism that appears independent from CTCF and insulation. The regulation of the interaction between PcG-bound regions by cohesin ultimately affects the expression patterns of Polycomb target genes [128].

Overall, PcGs play an unequivocal role in the 3D genome organization and in mediating the establishment of long-range interactions. Accumulating evidence suggests that functional role of Polycomb can be highly context-dependent, providing either a repressive or a permissive chromatin environment to regulate gene expression, particularly in the context of ESCs (for example in the context of bivalent domains, discussed in more detailed in **section 1.3.4**) [129, 111, 130].

1.1.5 Methods for profiling 3D genome architecture

The existence of euchromatin, heterochromatin and, more in general, of a highly organized chromatin structure within the nucleus in chromosome territories were first proposed in the nineteenth and twentieth centuries [1, 131, 132]. But it was only in the 1980s that the presence of chromosome territories was identified through fluorescence in-situ hybridization (FISH). FISH-based experiments confirmed that not only different chromosomes occupy a specific region within the nucleus, but also that highly actively transcribed

regions tend to associate together towards the center of the nucleus, while less transcriptionally active regions associated with the nuclear periphery [133, 134, 135]. A major leap forward in the understanding of chromatin organisation in 3D space occurred due to the establishment of chromosome conformation capture techniques (3C) and its derivatives, referred to as C-based techniques [136, 16] (**Figure 1.4**).

3C was first developed in the early 2000s and it allows analysis of the interactions between two chosen genomic loci through chromatin cross-linking (which captures protein-mediated or RNA-mediated interactions) and proximity ligation [16, 137]. In conventional 3C, existing contacts between two regions of interest ("one versus one" approach) are then identified by quantitative polymerase chain reaction (PCR) (**Figure 1.4**). 3C experiments, however, do have limitations: they require prior knowledge of the regions of interest and they are limited in the detection of longer range contacts [138, 139]. The need to overcome these limitations sparked the establishment of improved methodologies. Based on the same principles of 3C, other methods were developed shortly after in order to investigate the 3D conformation of regions of interest at a higher scale (defined as C-based techniques): circular chromosome conformation capture (4C) and 3C carbon copy (5C) [113].

4C, like 3C, is based on proximity ligation where interacting DNA fragments generate a circular DNA molecule. Here, the use of primers for a specific region of interest allows the identification of all the genome-wide contacts of the chosen viewpoint ("one versus many"). 4C, unlike 3C, does not require prior knowledge and allows the detection of contacts in *cis* (i.e. on the same chromosome) or in *trans* (i.e. on different chromosomes) (**Figure 1.4**). 4C was later combined with whole genome sequencing and it was applied to describe the dynamics of chromatin compartments during development for the first time. Furthermore, studies based on 4C methods gained first insights on the compartmentalization of the genome into active and inactive compartments [140].

5C allows the analysis of the spatial 3D conformation of larger genomic regions ("many versus many" approach). Using a forward and reverse primers, it makes use of a 3C library to amplify large genomic regions (in some cases they can be up to megabases long) offering greater resolution (**Figure 1.4**). However, the resolution is highly dependent on the design and availability of suitable primers for the region of interest. Nevertheless, 5C enabled the investigation of complex interactions of a given locus of interest and their *cis-trans* interaction networks, (for example, 5C-based studies helped uncover the organization of the Hox clusters, both in mouse and human [141]) and it provided first evidence

for the existence of TADs on the X chromosome [142]). However, it still doesn't allow the study of interactions on a genome-wide scale.

The development of Hi-C made it possible to identify contacts genome-wide ("all versus all"), both in *cis* and *trans*, simultaneously. Hi-C is based on the same principles of the other C-based techniques described above, but ligation is preceded by the biotin tagging of the DNA overhangs resulting from the digestion with restriction enzymes. Exploiting the streptavidin-biotin affinity, the final sample is enriched with ligation products containing fragments in close proximity in the nucleus. The final library is then sequenced via Illumina paired-end sequencing [18]. The newest development in recent years, Micro-C, sees the substitution of restriction enzyme digestion with micrococcal-nuclease (MNase), obtaining nucleosome-resolution level of chromatin folding [20] (**Figure 1.4**).

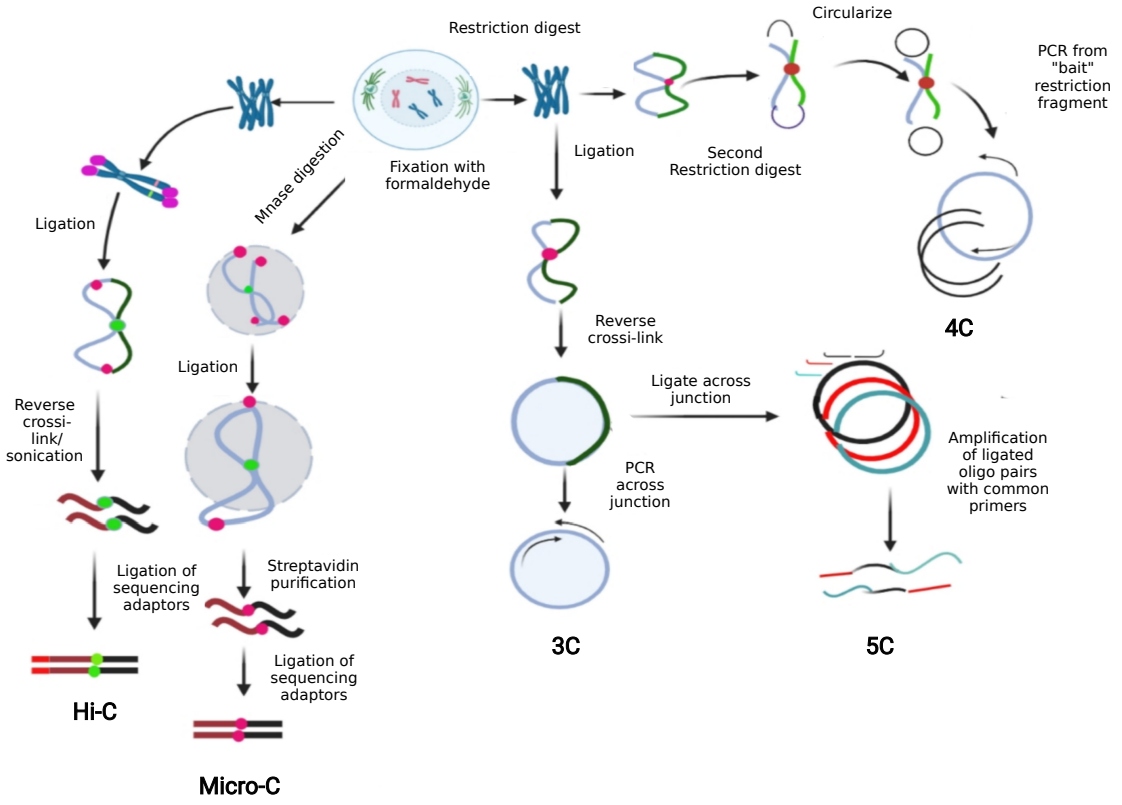


Figure 1.4: **Chromosome conformation capture (3C)-based methodologies.** All 3C-derived techniques involve isolation of nuclei and DNA, followed by formaldehyde cross-linking of the chromatin. For Hi-C, restriction enzyme fragments are labelled with biotin and, through streptavidin-biotin pulldown, the final library is enriched for fragments that are in close 3D proximity in the nucleus. In Micro-C restriction enzyme digestion is replaced by the digestion of chromatin using micrococcal nuclease (MNase) digestion, obtaining nucleosome-resolution levels. For 3C, ligation of restriction enzyme fragments is followed by Polymerase Chain Reaction across the junction while in 4C, restriction digestion and ligation are followed by a second restriction digestion to circularize the genetic material followed by PCR amplification using primer for the restriction fragment of interest. In 5C, reverse cross-linking is followed by the ligation of oligos across the junction which are then used as template for PCR-based amplification with a common primer. Figure adapted from Mohanta, T.K., et al., 2021 [113].

C-based techniques, particularly Hi-C/Micro-C, have driven most of the exploration of the 3D genome organization over the past decade and have allowed major advancement in our understanding of the genome hierarchical organization within the nucleus and its function [143]. In addition, the development of Hi-C has been key for elucidating the functional role of 3D organization of the genome in the fine regulation of transcription, which will be described in more details in **section 1.2**.

1.2 REGULATION OF TRANSCRIPTION

The DNA contains the genetic information necessary for the development of highly complex organisms. This includes protein-coding genes and non-coding regulatory elements that largely determine accurate gene transcription that is fundamental to determine cell identity and function [144].

1.2.1 DNA transcription: a brief overview

Transcription is mediated by RNA polymerases that use DNA as a template for the synthesis of RNA molecules. Three different RNA polymerases have been characterized in eukaryotes [145], which were later found to drive the transcription of different classes of genes: RNA polymerase I (RNAPolI), RNAPolII and RNAPolIII responsible for the transcription of large ribosomal RNA precursors, messanger RNAs (mRNAs) and long non-coding RNAs, transfer RNAs and small ribosomal RNAs, respectively [146].

In order for transcription to initiate, the RNA polymerase needs to gain access to the promoter region [147]. A promoter is usually defined as a DNA sequence typically located near the transcription start site (TSS) of a gene (i.e. in the proximity of the 5′ region) that is able to drive transcription through the binding of various factors (**Figure 1.5**). Active promoters are usually found in nucleosome depleted regions (NDRs) and they can be distinguished in different classes based on the DNA underlying sequence, the class of the gene transcribed and the type of RNAPol recruited. In the case of RNAPolII, bound promoters can contain CGIs or display the presence of a TATA element upstream of the TSS [148, 149, 150, 151, 152].

In general, RNA polymerases can not directly bind to promoters, instead their recognition depends on the binding of DNA-sequence specific transcription factors (TFs). TFs can act as a "bridge" to mediate the connection between polymerases and promoters and form specific pre-initiation complexes (PICs) on the promoter DNA [153, 154, 155, 156, 157] (**Figure 1.5**). In particular, in the case of RNAPolII the PIC is typically formed by TATA-box binding protein (TBP) and TFIIB which recruits the RNAPolII-TFIIF complex, forming a link between the polymerase enzyme and the promoter region [158, 159, 160, 161]. Typically, the main role of the PIC is to make the promoter DNA accessible. In the case of RNAPolII, this requires the recruitment of an additional factor, the DNA translocase XPB, a subunit of TFIIB that usually binds the DNA region downstream of RNAPolII

[162, 163].

The formation, stability and function of the PIC in mediating transcriptional initiation and progression are tightly regulated and will be the focus of the next section.

1.2.2 Transcription initiation, pausing and elongation

Transcription initiation by RNAPolII is mediated by its association with a co-activator complex known as Mediator [164] and with TFIIH complex which, through its helicase-associated activity, translocates the DNA towards RNAPolII and contributes to the formation of the transcriptional bubble. This promotes RNAPolII to initiate the synthesis of the complementary RNA molecules [165, 166, 167]. After the initial transcription of an RNA molecule between 20-60 nucleotides, pausing of RNAPolII can occur near the TSS of a subset of genes [168, 169, 170]. The pausing of transcriptionally enganged RNAPolII was initially observed at the promoter of the heat shock protein 60 (HSP60) in *Drosophila*, but it is now a known feature of both the *Drosophila* and the mammalian genome [171, 172]. RNAPolII pausing it is now considered a major mechanisms for the regulation of transcription and it can occur at promoters associated with a broad range of expression levels. However, it appears to be enriched at genes that require precise and prompt transcription, for example at promoters of genes involved in development. Several processes have been described to mediate the transition from paused RNAPolII to productive elongation of transcription [173, 174, 168].

RNAPolII contains a C-terminal domain (CTD) formed by a repeated heptapeptide sequence than contains several serine (S) residues which can undergo phosphorylation during transcription [175]. Upon initiation, phosphorylation of S5 mediated by CDK7 disrupts the PIC contact with Mediator and basal TFs and promotes the association with DSIF (5,6-dichloro-1-b- dribofuranosylbenzimidazole sensitivity -inducing factor) and NELF (negative elongation factor) instead, resulting in an inactive conformation of RNAPolII [176] (**Figure 1.5**). Typically, after capping of the 5′ of the nascent mRNA (necessary for efficient translation of the mRNA), pTEFb (positive transcription elongation factor b) mediates the phosphorylation of DSIF, NELF and S2 of RNAPolII CTD. Phosphorylated S2 allows the recruitment of the polymerase-associated factor (PAF) complex and TFIIS, causing the transition of RNAPolII from a paused to an elongation-competent conformation [177, 178, 176] (**Figure 1.5**). In addition to phosphorylation, paused RNAPolII is also characterized by the presence of other post-translational modifications, such as methylation and acetylation of the non-canonical K residues of the CTD, usually observed for

RNAPolII present at gene associated with low transcriptional activity [179, 180].

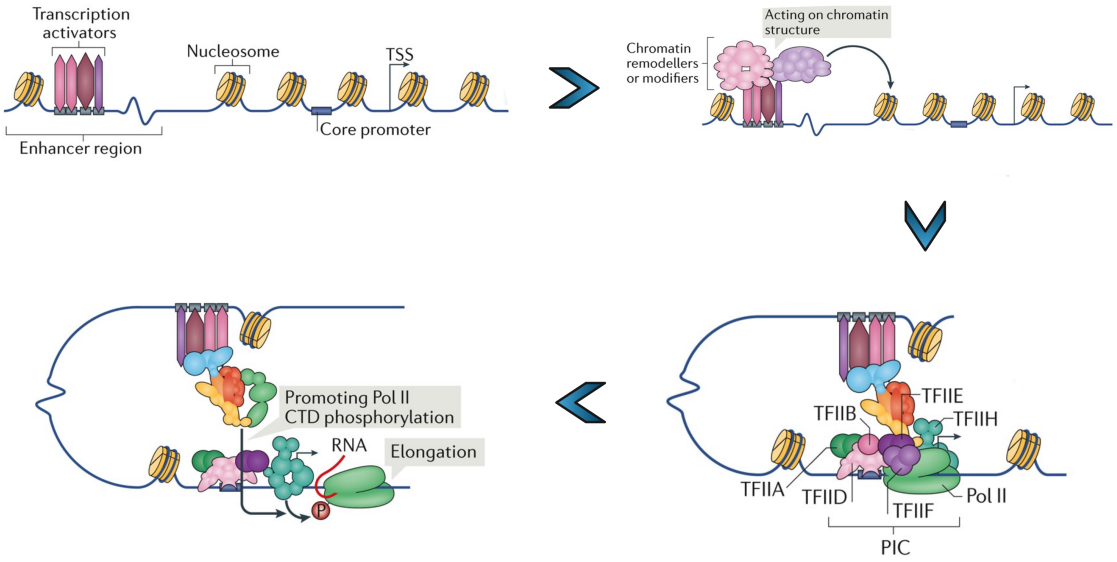


Figure 1.5: **Simplified schematic of DNA transcription by RNApolII.** Transcription activation starts with the binding of transcription factors (TFs) at the promoter region, usually defined as a DNA region in proximity to the transcription start site (TSS). RNA polymerases cannot directly bind to the promoter and their recruitment depends on the binding TFs, which mediate the connection between polymerases and promoters forming pre-initiation complexes (PICs). The PIC is assembled at the core promoter and it includes RNApolII and transcription initiation factor IIA (TFIIA), TFIIB, TFIID, TFIIE, TFIIF and TFIIH. The phosphorylation/de-phosphorylation of carboxy-terminal domain (CTD) of RNApoII at Ser5/Ser2 is necessary for RNApolII to switch between an inactive conformation and the elongation-competent conformation, respectively. Figure adapted from Soutourine, J., 2018 [181].

Another mechanism involved in RNAPolII pausing is the association with PcGs, which was initially observed in ESCs, but it has since been observed in differentiated and post-mitotic cells as well [182]. Typically, when RNAPolII is associated to PcG-bound regions it only displays phosphorylation of S5 and it is usually referred to as *poised* RNAPolII, to distinguish it from the "canonical" paused RNAPolII [183, 184]. Phosphorylation of S5 of *poised* RNAPolII at Polycomb-bound ESCs genes is mediated by ERK1/2 kinases and it has been shown that loss of ERK1/2 in ESCs leads to a loss of *poised* RNAPolII and dissoci-

ation of PRC2 at Polycomb-bound regions, further supporting the functional association between *poised* RNAPolII and PcGs [185, 186]. Particularly in ESCs, H2AK119Ub mediated by PRC1 plays an essential role in "holding in check" the *poised* RNAPolII at a subset of developmental genes (in particular, *poised* RNAPolII is found at a subset of bivalent domains which are described in more details in **section 1.3.4**) [183]. The interplay between RNAPolII poising and PcGs-mediated has been found to play a key role in the regulation of gene expression programs during development, perhaps supporting the idea of a dynamic equilibrium between an active and a repressed state of genes, especially in the context of ESCs [182].

1.2.3 Enhancers as transcriptional regulatory elements

Alongside promoters, non-coding DNA *cis*-regulatory elements (CREs) play an important role in the regulation of gene transcription. Amongst CREs, enhancers have been characterized to a greater extent. Evidence of the first enhancer element was found in the 1980s and since then, millions more have been identified in a myriad of different cell types. Due to advancements in available technologies, the pivotal role of enhancers in development and differentiation has become clear, as well as their role in disease progression through enrichment for non-coding variants, identified through population genetic studies [187, 188].

In the 1980s two parallel studies provided the first evidence for the existence of enhancers. Short DNA sequences able to drive the expression of a gene were identified in the simian virus 40 (SV40) [189] and in the genome of the sea urchin [190]. These sequences did not have distinctive promoter features and were located distant from the promoter of the gene. These observations were later confirmed by several studies that showed the ability, indeed, of a SV40 72bp DNA sequence to drive gene expression using a reporter system in mammalian cells [191, 192, 193, 194]. In addition, it was shown that this sequence was able to activate β-globin, both mouse and human, regardless of its distance from the gene promoter or orientation [191, 192]. Independence from distance and orientation have since been recognised as common characteristics of enhancer elements.

These findings fueled the identification of more regions with enhancer-like features and enhancers were soon found in multiple organisms, from bacteria and yeast [195] to *Drosophila* [196] and mammals [197, 198, 199]. In the mammalian genome in particular, enhancers outnumber by far gene coding regions and they can be located at any distance from their target genes or within them, as well as being placed within an unrelated gene

[200, 201, 202].

Understanding how enhancers are able to exert their function is still an area of active research. A multi-layered mechanism is likely to be behind enhancers ability to regulate their target genes with a high degree of specificity. At the level of DNA sequence, enhancers can be bound by sequence and cell-type specific TFs and co-factors involved in the recruiting of the transcriptional machinery at promoters. At a higher-order level, chromatin architecture and the folding in 3D of the genome through chromosome loops and, perhaps, phase-separation, also play a crucial role in allowing enhancers to accurately orchestrate gene expression programs [187, 188].

1.2.4 Enhancers as landing sites for transcription factors (TFs)

As more enhancer-like sequences were identified, it became evident that these regions display a high density of motifs recognized by TFs [203]. This led to the emergence of a model whereby enhancers represent landing sites for TFs which collectively increase the activity of promoters and provided the first explanation as to how a specific combination of TFs may regulate gene expression in a cell-type and in a condition-specific manner [204, 205]. However, due to the highly packaged conformation of chromatin, some required enhancer elements could potentially be found tightly wrapped around nucleosomes, and this would therefore hinder the recruitment of the relevant TFs. To compensate for this, a number of cell-type specific TFs, known as pioneer factors, are able to bind to their consesus motif and displace nucleosomes either by exerting chromatin remodelling functions themselves or via the recruitment of different chromatin remodellers [206, 207]. Numerous studies have shown that pioneer TFs play an important role, particularly during development. An example is given by the Zinc Finger Early Drosophila Activator (Zelda) in *Drosophila* and its role in making enhancers accessible to allow zygotic gene activation (ZGA) [208, 209, 210]. Examples of pioneering TFs have also been identified in mammals, both mouse and human, such as FOXA1, OCT4, SOX and KLF4, which play an essential in the control of pluripotency [211], and ASCL1, PAX7, PU.1, GATA4 and P53 [212]. However, the binding of pioneer TFs alone is often not enough to lead to enhancer and promoter activity and they rely on the cooperation of additional TFs that respond to specific environmental cues and are able to bind to the accessible enhancers. It is, ultimately, the cooperative binding of multiple TFs that make the enhancer regions fully accessible and able to stimulate gene expression [204, 213, 187].

TFs that bind enhancers in a sequence-specific manner are also able to recruit factors involved in the assembly of the transcriptional machinery, in addition to co-factors which promote chromatin remodelling and the deposition of histone modifications [206]. Well known examples of co-factors are the FACT and the SWI/SNF complex to facilitate chromatin remodeling. Factors like Mediator, p300 and CREB-binding protein (CBP) are typically associated with promoting enhancer activity together with Bromodomain containing (BRD) proteins, such as BRD4 [214, 215, 216, 217, 218]. In particular, Mediator and BRD proteins have also been associated with so called "super-enhancers", defined as regions that span over longer DNA loci with regulatory activity and have been shown to be able to drive gene expression at a greater extent compared with regular enhancers [219, 220, 221, 222].

It is believed that the organization of the TFs-specific binding motifs could affect the specificity and efficacy of a given enhancer. The principles that this organization follows - typically referred to as enhancer "grammar" - are still an area of active investigation [205, 188]. Nevertheless, several models have been proposed to describe the lexicon of TFs motifs at enhancer regions (**Figure 1.6**). The "enhanceosome" is the most rigid model and it calls for strict motif organization, positioning and spacing. It requires the recognized motifs to remain in the same order and orientation for the enhancers to work. Reported examples of "enhanceosome-like" structures are rare, one being the viral inducible interferon-β (IFN-β) enhancer in mammals that requires the binding of eight TFs [223, 224]. On the more flexible end, the "billboard" model has been proposed, where the presence of a given TF binding motif has a stronger contribution than its positioning and orientation, perhaps implying a lower degree of cooperativity between TFs for DNA binding. This type of enhancers are likely to be widespread in vertebrates and they tend to contain sub-optimal binding sequences that may allow for a more rapid motif turnover and, ultimately, play a role in determining enhancer specificity [225, 226, 227, 228]. Somewhere in between these rigid and flexible models, there are a large number of enhancer regions that lack specific underlying consesus motifs, where TFs recruitment happens as a combination of DNA binding and protein-protein interaction between TFs themselves [204, 229]. This intermediate model of enhancer grammar is known as "transcription factor collective" model and was first postulated in *Drosophila*, where a set of enhancers involved in heart development, despite binding to the same set of TFs, didn't share any similarity at a DNA sequence level [229].

It is thought that the vast majority of enhancers are likely to lie on a spectrum between

the ”enhanceosome” model and the ”billboard” model, with strictly defined orientation and positioning required for some enhancers, but not others [205].

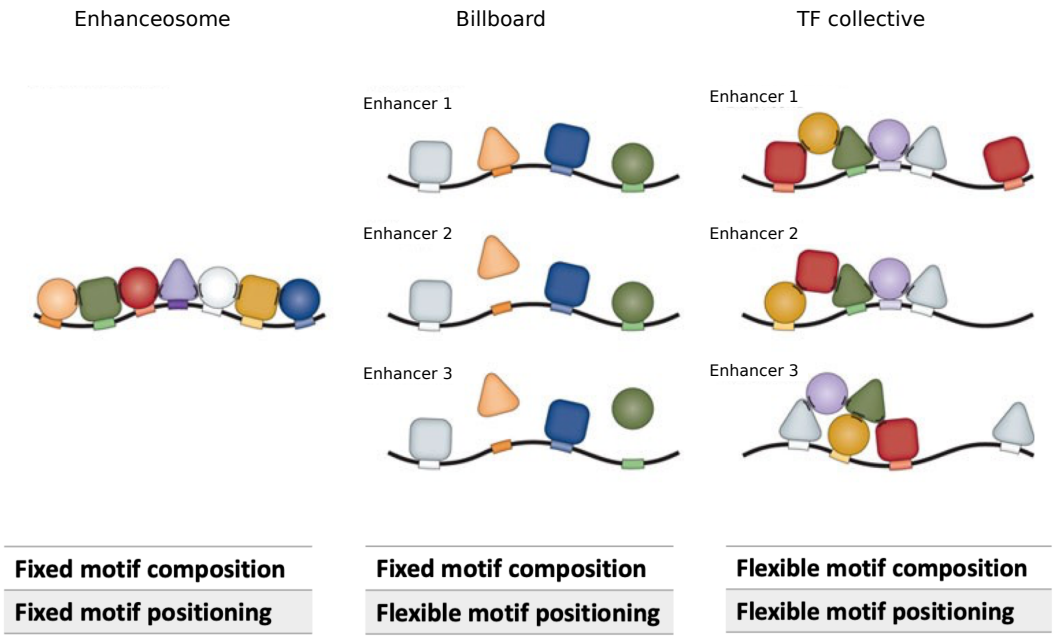


Figure 1.6: **Current models of enhancer ”grammar”. Enhanceosome**: the binding of all transcription factors (TFs) at the enhancer regions is essential for the activation of the enhancer. It is characterized by a strict DNA motif composition and positioning (motif grammar). **Billboard**: the positioning of TF binding sites is flexible, while the presence of a given TF motif has a stronger contribution to the enhancer activity. **Transcription factor (TF) collective**: the same set of TFs bind to many enhancers (as an example, only five are illustrated). The presence of a given TF motif and its positioning are subject to greater flexibility and TFs can occupy each enhancers in a different manner. According to this model, the recruitment of TFs to the enhancer can be mediated via protein–protein interactions. Figure adapted from Spitz, F. & Furlong, E., 2012 [204].

Nevertheless, the different enhancer architectures will have different implications for how well the enhancer regulatory landscape is conserved. For example, in the case of the ”enhanceosome” model a single mutation would be expected to greatly affect the enhancer function. As a result, this type of enhancers tends to be highly conserved across evolution, presumably because of their involvement in the regulation of genes that follow much stricter transcriptional programs [204]. These highly conserved elements tend to cluster near genes involved in the regulation of developmental transcriptional programs

and in some cases, their mutation can lead to disease in human, as seen for the mutation in the *Shh* enhancers that results in polydactyly [230, 204]. However, despite there being some well characterised enhancers that fall in this category, they are in the minority and it is much more common to observe *de novo* emergence or loss of enhancers. Even amongst enhancers retaining conserved activity, the conservation signature is, on the contrary, often very weak [231, 232]. For example, poised enhancers (described in more detail in **section 1.3.4**) that are involved in the regulation of developmentally important genes do so despite them not being well conserved at the DNA level across species, hinting to a much more flexible organizational model [233]. *Even-skipped* stripe 2 enhancer (*S2E*) represents a well characterized example demonstrating such flexibility. In particular, *S2E* controls the expression of *even-skipped* stripe 2 expression in *Drosophila*, a gene encoding for a homeo-domain-containing protein (Eve) required for the development of parasegments, both odd- and even-numbered. Despite the functional conservation and the expression pattern of this gene being strongly conserved, its enhancer element *S2E* has undergone a major motifs re-shuffling across species, with changes in the binding-site sequences and their spacing [234, 235].

Overall, given that it is common to observe the loss or gain of TF-binding motifs within enhancers without there being any discernible impact on enhancer function, this would support a model for more flexible enhancer grammar being utilised throughout evolution.

1.2.5 Enhancers' chromatin signature

Enhancer regions typically display specific chromatin features. Active enhancers, in particular, are usually characterized by a low nucleosome occupancy allowing for a more open and accessible chromatin environment. The depletion of nucleosomes at active enhancers allowed the use of technique such as DNase-I hypersensitivity sites sequencing (DNase-seq) and assays of transposable-accessible chromatin via sequencing (ATAC-seq) to identify Nucleosome-Depleted Regions (NDRs) with putative enhancer activity [236, 237]. Chromatin accessibility, however, is not only associated with enhancers, but also with: promoters, regions that are "primed" for gene activation despite not being active *per se* and regions bound by structural proteins such as CTCF [238, 239, 240, 241].

PTMs of histone tails represent another suitable readout to define enhancer regions and provide information on their activity state. Nucleosomes at enhancer regions can, indeed, carry specific PTMs including H3K4me1 and H3K27ac which, in combination, are considered a hallmark of enhancer activity and are often used to annotate active enhancers

a priori [242, 243, 233, 244, 245]. H3K27ac can also be found at promoters, but they can be distinguished from enhancers based on the higher levels of H3K4me3 instead of H3K4me1 [240, 188].

The methylation of H3K4 is mediated in mammalian cells by MLL/Set1 family of methyltransferases [246, 247]. In general, Set1a and Set1b together with MLL1/2 are responsible for laying H3K4me3 at promoters [248, 249] or it can be found at most housekeeping enhancers, while MLL3/4 are mainly involved in the deposition of H3K4me1 at enhancers, specially at developmental enhancers, as it has been observed in various systems (e.g. adipogenesis, cardiac development, lymphogenesis) [250, 251, 252, 253, 254, 255]. According to a number of studies, H3K4me1 is usually found upstream of H3K27ac at enhancers and, in some cases, enhancers can display H3K4me1 but be devoid of H3K27ac in a state that has been described as "primed", usually not associated with detectable transcription [233, 243, 245, 244, 218, 256]. Active enhancers recruit p300 and CBP, which to date are the only known HATs able to catalyze H3K27ac [257]. H3K27ac is conventionally used to identify regions with putative enhancer activity and, although it represents a powerful predictive mark, its precise functional role in the regulation of gene transcription still remains unclear. Both promoters and enhancers can display acetylation on a variety of lysine residues, such as acetylation of lysine 16 of histone H4 (H4K16ac), of lysine 122 of histone H3 (H3K122ac) or of lysine 64 of histone H3 (H3K64ac), associated with enhancer or promoter activity [258, 259, 260]. It remains unclear as to which specific acetylated residues, or combination of thereof, serves a specific purpose. It is possible that the recruitment of specific HATs, hence the presence of a specific combination of acetylated residues, may contribute to the binding of specific TFs and the establishment of cell-specific transcription patterns [261]. Recent studies suggest that, despite its correlation with trascriptional activity, loss of acetylation at H3K27 has little to no observable consequence and that H3K27ac may actually be dispensable for the recruitment of TFs and the transcriptional machinery at both enhancers and promoters [262, 263]. As mentioned above, it is known that numerous histones' lysine residues can be acetylated (for example H3K9ac, H3K18ac, H3K23ac, H3K122ac) and these modifications often appear to be co-enriched. Therefore, it is plausible to speculate that the synergistic action of different acetylated residues, rather than individual events, is more crucial for transcription activity [262].

In recent years, studies in mESCs and hESCs have shown that, beyond the active and primed state, a large group of enhancers are marked by H3K4me1, but they dis-

play trimethylation of H3K27 (H3K27me3), rather than H3K27ac [233]. As previously described, H3K27me3 is deposited by PcGs and it is usually associated with repression of transcription. These elements represent a group of enhancers termed as "poised" and are generally located near genes that play key roles during development (poised enhancers and their role in development is described in **section 1.3.4** [233, 264, 265].)

The set of histone modifications described above is routinely used to identify enhancers in a genome-wide manner, with the possibility to distinguish enhancers in their active, primed or poised state. Although they represent only a fraction of the full ensemble of possible chromatin marks, identifying these shared features made it possible to annotate enhancers in cell-type specific contexts and independently of their conservation status [261]. Untangling their functional role, on the other hand, has been proven difficult, mainly due to technical limitations. Nevertheless, it is becoming clear that the role of PTMs goes beyond being mere signposts of the state of an enhancer. They actively participate in the recruitment of context-dependent TFs and determine enhancers' chromatin accessibility [266, 267, 268, 269]. In some instances, histone modifications have been found to contribute to "poise" enhancers for future use and they may have a role in mediating the communication between enhancers and promoters in the 3D space [270, 264].

The field is gradually moving away from the on/off dichotomy as more evidence appears of poised and pioneer enhancers and their involvement in a finer regulation of gene expression patterns.

1.2.6 Enhancer-promoter crosstalk

How enhancers can activate gene transcription is a question that the field of eukaryotic gene regulation is actively trying to answer. One of the main features of enhancers is in their ability to regulate the transcription of promoters over large distances, an example of which is the *Shh* ZRS enhancer which is situated one mega base (Mb) away from its target gene [230]. Therefore it is likely that to exert their function enhancers must somehow communicate with promoters [188].

As discussed previously, 3C based techniques and, particularly Hi-C, have been key in the identification of different layers of 3D genome organization. However, due to the high complexity of Hi-C data, detailed analysis of interactions between individual loci, such as enhancer-promoter interactions, is not possible without hitting limitations of resolution and statistical power. To overcome these limitations, Capture C and Capture Hi-C

were developed, whereby a sequence enrichment approach is coupled with 3C or Hi-C experiments. Here interactions containing regions of interest are captured via sequence hybridization, thereby reducing the Hi-C library complexity and gaining much greater resolution [271, 272, 273]. One of the first examples of such strategy is Promoter Capture Hi-C (PCHi-C), first developed in the early 2010s [274, 275, 276]. Here Hi-C libraries are hybridized with RNA tagged with biotin (also referred to as *baits*) designed to be complementary to all annotated promoter elements, resulting in the enrichment for interacting products that contain promoters at one end and their interacting *other-ends*, which in most cases are enriched for regulatory elements. PCHi-C-based studies have shed lights on the promoter-enhancer interaction network and its rewiring in different contexts and cells types (e.g. during differentiation in both mouse and human embryonic stem cells (ESCs) [277, 278], in adypocytes [279] and keratinocytes [280], upon loss of cohesin and its role in mediating loop extrusion [281, 282, 283, 284, 285, 286]). However, baits can be designed to capture any regions of interest other than promoters. For example, several studies have customized baits to capture disease-associated risk loci identified through Genome-Wide Association Studies (GWAS) with the aim to link non-coding variants to their target genes exploring their 3D chromosomal interactions [287, 288]. Moreover, more recently sequence-capture approaches have been coupled with Micro-C (Micro-Capture-C) which allows to look at interactions of regions of interest reaching single-base pair resolution [20, 289].

Several models have emerged that aim to provide an explanation for enhancer-promoter communication [188] (**Figure 1.7**). In the ”looping” model the DNA string that separates the enhancers from their targets gets ”looped out”, bringing the two elements in close proximity in 3D space [290, 291]. The alternative ”linking” model envisions that proteins bound to the enhancers recruit a series of additional proteins that ultimately serve as a bridge to link the enhancer with its promoter [292, 293, 294]. Similarly to the ”linking” model, the ”sliding” or ”tracking” model also suggests that proteins are involved in deploying the transcriptional signals from the enhancer to the promoter by scanning the chromatin that separates them [295]. For both the linking and the sliding models, the protein complexes would eventually become hindered by the transcriptional machinery, perhaps as another mechanism to regulate transcriptional output or to prevent inappropriate gene expression [295, 296, 297]. While all three models proposed unique and distinct mechanisms, in practice it had been difficult to distinguish between them. However, with

the emergence of novel techniques, it has become possible to gather overwhelming experimental support for enhancers that physically contact their cognate promoters in order to achieve gene expression. A seminal study proved, indeed, that inducing contact between the promoter of the β-globin gene (*Hbb*) and its locus control region (LCR) enhancer led to induction of *Hbb* transcription [298, 299]. However, recent studies in *Drosophila* using Hi-M, an imaging based technology that enables the study of chromatin organization and transcriptional status in intact embryos, showed that chromatin loops between enhancers and their targets can precede the formation of TADs or the detection of transcriptional activity in early development, suggesting that 3D chromatin loops may not always have a regulatory function [300].

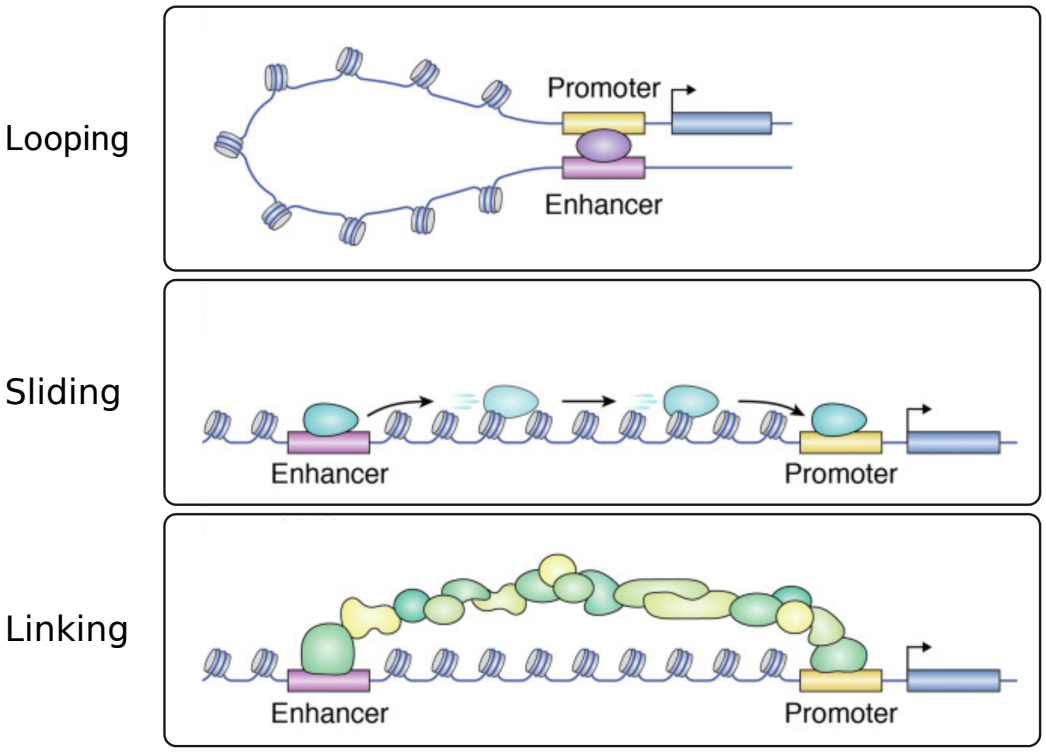


Figure 1.7: **Schematic of current models of enhancer-promoter communication. Looping:** The DNA between the enhancer and its target promoter is "looped out" in the process of bringing the two regions in close proximity in 3D space. **Sliding:** a protein initially associates with the enhancer region and deploys the activating signal by scanning the DNA region that seprates the two elements. **Linking:** a series of transcription factors (TFs) are recruited at the enhancer region and contribute to the formation of a protein bridge to transmit the activating signals from the enhancer to the promoter. Figure adapted from Popay, T. & Dixon, J.R., 2022 [301].

Generally, enhancers represent dense clusters of binding motifs which recruit TFs re-

sponsible for the transcription of a specific target promoter [295, 302]. It could be speculated that such specificity may also be the result of direct enhancer-promoter contact. 3-C derived techniques and FISH-based assays uncovered important features of the enhancer-promoter crosstalk [303]. They confirmed that, indeed, the model on enhancer-promoter interaction via looping is widespread in many cases, although a more complex picture is slowly emerging [303].

There are many ways in which enhancers can control gene expression, including: chromatin reorganization, recruitment of transcriptional machinery (including RNAPolII), removal of repressors and/or facilitating pause-release of RNAPolII [304]. However, a clear mechanism to explain how transcriptional activation is achieved is still lacking. Hypothetically, transcription activation mediated by enhancers could be achieved in two main steps: first, TFs are recruited [305, 306] and, based on their ability to establish protein-protein interactions, they then recruit secondary TFs, activator and co-activators (e.g. Mediator, p300, CBP) [307, 306]. These considerations suggest the possibility that TFs-bound enhancers regulate gene transcription not by actively participating to the transcriptional activation *per se*, but rather representing platforms for the recruitment of additional factors that ultimately lead to activation of transcription [303].

The lack of a defined structure in the nucleus would make enhancer-promoter communication harder to achieve. In most cases, it is now evident that clusters of CREs and target promoters, along with by-stander genes, are spatially constrained in evolutionary conserved "blocks", or TADs [308, 309]. Enhancers and their target genes almost always reside within the same TAD and, in some cases, perturbing TADs' boundaries can result in a "reshuffle", whereby enhancers can contact and affect the expression of genes that would otherwise reside in an inaccessible TAD [310, 311]. However, there are cases where enhancers and promoters can contact each other across TAD boundaries. Nevertheless, there is evidence supporting the role of CTCF in mediating chromatin-loop formation in cooperation with cohesin and, in some cases, acting as an insulator preventing aberrant chromatin loops formation (reviewed in [188]). However, this is not always the case as the effects of structural variations on gene expression can be highly context-dependent [312]. Indeed, studies in *Drosophila* showed that TAD-boundary removal or TAD rearrangements often do not result in major effects on gene expression [313, 314]. Furthermore, it has also been shown that upon deletion of both CTCF and cohesin there is a very small impact on gene expression and evidence are starting to emerge implying that,

perhaps, deletion of CTCF might only impact longer-range interactions (i.e. > 100kb), suggesting that other mechanisms may be at play [315].

Although major players, CTCF and cohesin are not the only factors involved in the establishment of chromatin loops. There is now increasing evidence of PcGs mediated looping (as discussed in **section 1.1.4**) which are likely independent of CTCF and cohesin and, in some cases, might counteract their function [108, 316, 317]. YY1 and ZNF143 are two zinc-fingers which have recently been found to be responsible for the establishment of many lineage-specific enhancer-promoter loops. On the contrary to what has been seen for CTCF and cohesin, their depletion leads to a significant decrease in the observed loops, greatly affecting gene expression levels [318, 319, 320, 321, 322]. LIM domain-binding protein 1, LDB1, has been implicated in the formation of chromatin loops via protein-protein interactions [323]. It was first described as a looping factor for the β-globin gene in mouse and it has since been implicated in facilitating chromatin interactions during differentiation [324, 325]. Like in the case of YY1 and ZNF143, depletion of LDB1 also leads to disruption of gene expression, but contrary to YY1 and ZNF143 mediated loops, LDB1 looping seem to be independent of CTCF and cohesin [325, 326, 327, 328, 329]. There is also evidence that transcription itself can stabilize or disrupt enhancer-promoter contacts [330] as in the case of *eve* locus in *Drosophila* [331].

Looping may not be the sole mechanism through which enhancers exert their regulatory function. An example of a looping-independent mechanism is given by a proximal enhancer element that controls the transcription of the *Shh* gene. Indeed, 3-C derived techniques have shown that, upon transcriptional activation of *Shh* the enhancer moves away from the gene (a phenomenon observed even when artificially recruiting activator such as VP64 and Mediator) [332]. An additional example in support of a looping-independent mechanisms is given by the case of *Sox2*. Live-cell imaging data show no association between transcription of the *Sox2* gene and the distance from its enhancer, bringing forward more evidence for an alternative mechanism that doesn't involve proximity or direct contact of enhancers and promoters [333]. Furthermore, recent studies based on Hi-C or Micro-C, have observed that during *Drosophila* development, multiple enhancers don't seem to contact their target promoter, irrespective of their activity status [334].

One hypothesis that it has recently gained ground is the ability of chromatin to form membrane-free condensates via liquid-liquid phase separation (LLPS). It is known that various biochemical reactions undergo LLPS and form condensates within the cell where

molecular processes can efficiently take place without requiring membranous boundaries [335, 336]. Two factors play an important role in phase separation: a concentration of molecules high enough to trigger separation into two different phases and molecules which conductivity ensure the continuous interactions between them [337]. LLPS can be mediated by intrinsically disordered regions (IDRs) of interacting proteins, but also by the modular structure of RNA or high density of TFs as it occurs, for examples, at enhancer regions (reviewed in [303]). Enhancer hubs, for instance, are described as aggregates where multiple enhancers contact a single promoter or, vice versa, a single enhancer can contact multiple promoters, forming a highly interacting regulatory network. This inevitably results in a high density of TFs leading to the recruitment of high concentrations of activators, co-activators, components of the transcriptional machinery and related factors. Histone proteins themselves have the intrinsic capability of triggering LLPS, together with Mediator, BRD4 and CTCF [338, 339, 340].

Although the role of LLPS in mediating regulation of gene expression hasn't been systematically assessed, it offers an alternative explanation to enhancer-promoter communication. LLPS could, perhaps, provide a different mechanism for cases where enhancers regulate in a "contact-less" manner the transcription of their target promoters. Additionally, the plasticity and reversibility of the nature of LLPS could represent a reliable mechanism to ensure correct expression patterns of genes that are co-regulated simultaneously [341].

Undoubtedly, enhancers bear the information necessary for correct spatio-temporal gene expression. The development of novel methodologies has allowed to gain initial insight of the mechanisms through which enhancers communicate with their target promoters, painting a complex and multilayered picture of the enhancer-promoter crosstalk. Overwhelming evidence has been gathered over the years showing how enhancers "loop-over" their target genes, physically contacting them. However, recent studies have started to uncover mechanisms that go beyond looping. The continuous advancement of the available technologies is allowing a more direct investigation of non-looping mechanisms and, more in general, enhancer regulatory logic.

1.3 DYNAMIC EPIGENOME CHANGES BETWEEN PLURIPOTENCY STATES

Pluripotency describes the capacity of cells to give rise to all three embryonic germ layers and primordial germ cells (PGCs), with the exception of extra-embryonic tissues. Although it represents a transient state *in vivo*, it can now be derived *in vitro* when providing cells with artificial external cues to help keep them in a self-renewal state [342, 343]. Pluripotency is a highly dynamic state that presents itself in different nuances at different stages of pre- and post-implantation development [344]. According to their source of origin, different types of pluripotent stem cells can be isolated in vertebrates, including humans [345].

1.3.1 Pre-implantation development in human: brief overview

Embryonic development begins with the fertilization of the oocyte by the sperm to form a diploid zygote. From that moment, the zygote undergoes a series of symmetrical mitotic division that lead to the formation of a 2-cell and a 4-cell stage embryo within the first two days after fertilization. During this initial stage, the embryo is still transcriptionally silent. Embryonic genome activation then occurs in two waves: a first minor wave is observed at the 4-cell stage and a second, bigger one at the 8-cell stage, within $\sim$3 days post fertilization [346, 347]. In general, the 8-cell stage is when essential morphological features and epigenetics remodelling occur. Between the 8-cell stage and the 16-cell stage the embryo undergoes compaction and it forms a tightly packed sphere known as morula and after five days post-fertilization the human embryo reaches the blastocyst stage and the inner cell mass (ICM) is formed [348, 349, 350]. During its maturation, the blastocyst expands into forming the trophoblast and the ICM divides into two lineages: hypoblast and epiblast. Trophoblast cells will then go ahead to form the placenta, the pluripotent epiblast progenitor cells will form the embryo proper and the hypoblast (or primitive endoderm) will determine the formation of the yolk sack as development proceeds [351, 352, 353]. Finally, between seven to ten days after fertilization, the implantation takes place (**Figure 1.8**) [354].

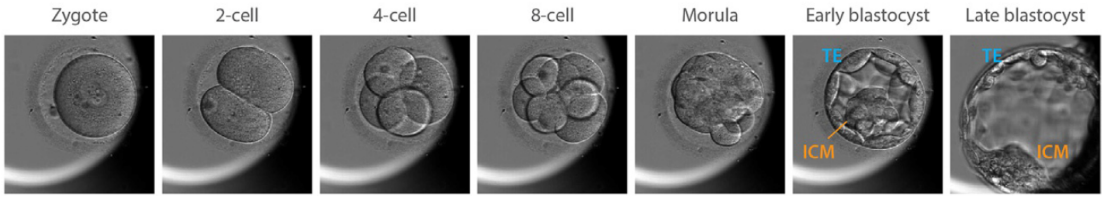


Figure 1.8: **Embryonic progression from one-cell zygote to late blastocyst.** Following fertilization, the zygote undergoes a series of mitotic divisions, forming the 2-cell and 4-cell stage embryo. Cavitation determines the formation of a blastocyst, consisting of: inner cell mass (ICM) and trophectoderm (TE). The ICM further segregates into a pluripotent epiblast and a primitive endoderm layer. The ICM further divides in hypoblast and epiblast which will go ahead and generate the embryo proper and the yolk sack, respectively. Figure adapted from Wamaitha, S.E. & Niakan, K.K., 2018 [355].

Human implantation cannot be observed *in vivo*, therefore most of the current knowledge comes from studies on the Carnagie histological sample series and is based on comparative analysis with primates and other model organisms [356, 357]. With the advent of *In Vitro* Fertilization (IVF) and Assisted Reproductive Technology (ART), which provided access to a source of pre-implantation human embryos, our understanding of human pre-implantation embryo development has increased [358]. Indeed, human ES-like cells were first isolated from human embryos cultured in the presence of a human oviduct feeder layer and capable to form the ICM, from which cells could then be isolated. ICM-isolated cells cultured in presence of mouse embryonic fibroblast gave rise to the first stable human ESCs lines which propelled the investigation of early development in the human context [359, 360, 361, 362].

The study of human pre- and post-implantation stages presents many technical and legal limitations. The following section will focus on the advancements of *in vitro* culture systems, providing an important resource to shed light on the initial stages of early embryogenesis.

1.3.2 Naïve and primed pluripotency

First reports of mESCs being isolated from the ICM of late blastocysts were given in the early 1980s [363, 364]. These cells could contribute to the formation of chimeras after injection in the blastocysts and, for many years, this became the gold standard test for pluripotency [365]. Almost two decades later, hESCs were isolated for the first time, but

it soon became clear that mESCs and hESCs required different culture conditions. Indeed, when mESCs were cultured under conditions specific for hESCs a new type of stem cell could be isolated, namely EpiSCs [366, 367, 368, 369]. Mouse EpiSCs (mEpiSCs) are more similar to hESCs and show similar characteristics: X-chromosome inactivation (Xi), poor survival when resuspended in a single-cell state and failure to contribute to the formation of chimeras, implying a more "restricted" pluripotent capacity of these cells [370, 371].

Based on these findings two different state of pluripotency were suggested: **naïve** and **primed** [372]. Naïve mESCs are cells derived from ICM of pre-implantation blastocyst, typically cultured in serum/LIF or 2i/LIF condition (i.e. leukemia inhibitory factor LIF + two additional inhibitors for MEK and GSK3) and they usually display "dome-shaped" colonies [345]. Primed mEpiSCs, instead, are derived from post-implantation ICM and they usually require the presence of Activin and FGF signaling. They typically display monolayered colonies, morphologically more similar to hESCs [370, 371].

Despite some obvious similarities, conventional primed hESCs are not identical to mEpiSCs. Although they both display post-implantation characteristics, the exact positioning of hESCs on the developmental time line is still uncertain due to both a higher degree of heterogeneity of the cell lines available and the lack of a real human reference for early post-implantation embryogenesis. Given the major differences between mouse and human early embryogenesis, in recent years a lot of effort has gone into the development of culture conditions that can stably sustain hESCs and, more specifically, to derive the naïve state in humans. Different ways to derive naïve cells are now available: 1)they can be be generated by resetting conventional iPSCs [373] and 2)they can be derived directly from dissociated human ICM [374]. These approaches all generate cells that display many features that typically distinguish the pre-implantation epiblast [375, 376], as well as the expression of factors typical of the naïve state, such as KLF4, KLF17 and TFCP2L1 [374, 377].

Generally, the transitioning from a pre-implantation to post-implantation is a period of major changes and it is crucial for the cells to acquire differentiation competence during a process of "capacitation".

Epigenetic factors undergo striking re-organization during this transition (**Figure 1.9**) [370].

DNA hypomethylation is a specific hallmark of naïve pluripotency. It has been hypothesized that the transient loss of DNA methylation that occurs during early embryogenesis

may contribute to reset the epigenome, granting the right level of plasticity to achieve proper cell differentiation. However, it is important to note that the levels of DNA methylation observed can vary across different naïve cell lines and it strongly depends on their culture conditions. Some cell lines can show 70% of DNA methylation (similar to levels observed in primed cells), whereas some naïve hPSCs maintained in different conditions display DNA methylation levels close to the ones observed in the human blastocyst, set between 20%-30% (reviewed in [378]). Nevertheless, naïve and primed hPSCs show substantial differences in the distribution of the methylation status across the genome, specially at CGIs. Imprinted gene control regions, for example, are typically methylated in the embryo, but show lower methylation levels in naïve cells [379, 380]. However, the difference observed might be a consequence of the continuous inhibition of MEK, typical of the naïve mPSCs culture conditions, contributing to the suppression of the DNA methylation machinery [381]. Despite the controversial evidence (some MEK-dependent conditions in hPSCs don't lead to loss of DNA methylation), this seems to suggest that the right level of MEK inhibition or the presence of other factors might contribute to the regulation of DNA methylation levels in pluripotent cells [378].

Xi is an additional epigenetic feature that allows to discriminate between naïve and primed hPSCs, in female cell lines. Conventional female primed hPSCs exhibit one inactive copy of X-chromosome, while naïve hPSCs display the presence of two active copies, an aspect that recapitulates what is observed in the pre-implantation embryo [378].

Lastly, histone modifications show distinct patterns between naïve and primed hPSCs [370]. Despite the inherent difficulty to assess differences for every histone modification, changes occurring for the most critical ones have been described and they remain a focus of current research. Particularly, H3K27me3 levels have been reported to be notably lower in naïve hPSCs compared to their primed counterpart and it has been an accepted criterion to discern naïve cells from primed (reviewed in [382]). However, recent studies showed that H3K27me3 absolute levels are extremely abundant in naïve cells, and rather than being absent, H3K27me3 shows a distinctive broad coverage of the genome. Therefore, it is likely that when cells transition to their primed state there is no real acquisition of new peaks of H3K27me3 *per se*, but rather a "sharpening" of the existing "background" of this particular mark. It was also recently shown that H3K4me3 is scarcely detectable when cells are found in their naïve state and it might only accumulate when transitioning to the primed state, affecting the ability to detect bivalency in naïve cells [383].

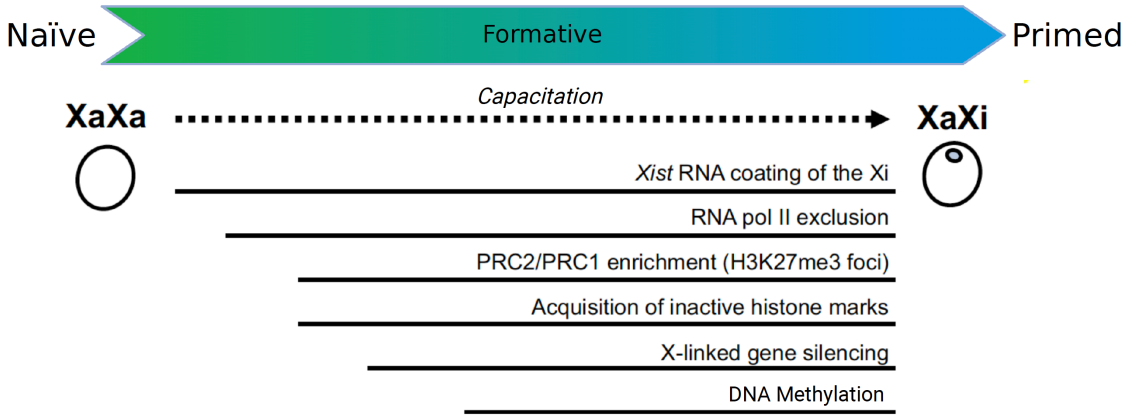


Figure 1.9: **Outline of the epigenetic changes between naïve and primed hESCs.** As hESCs transition from the naïve to the primed state of pluripotency, a striking re-organization of epigenetic factors takes place, including: X inactivation mediated by Xist; DNA-methylation; PRC1/PRC2 and H3K27me3 re-organization. Figure adapted from Takahashi,S., et al., 2017 [384].

In general, the naïve-to-primed transition represents a valuable model to look into processes that occur as early as the second week of gestation *in utero* and that would, otherwise, be inaccessible.

1.3.3 Epigenome rewiring in early embryogenesis

As pluripotent cells transition from the naïve to the primed state they undergo a major reorganization of their epigenome. These epigenetic changes occur in parallel with re-organization of the gene regulatory network and can result in different enhancer usage between the two pluripotency states [385]. Indeed, a process described as enhancer decomissioning has been observed between naïve and primed cells, whereby enhancers that are active in the naïve state show gradual loss of H3K27ac and are eventually repressed in the primed pluripotency state [386, 387]. An additional example of different enhancer usage is given by the distal enhancer of the *OCT4* that is preferentially utilized in the naïve state, to then rewire to its proximal enhancer specifically in the primed state [370, 388]. This implies a re-organization of long-range chromatin interactions and of the 3D genome organization between the two states of pluripotency.

Hi-C and CHi-C studies in hPSCs show that, overall, TAD structures are largely preserved between the naïve and primed states, although recent studies show that, in some cases, TADs boundaries in naïve cells can be expanded or contracted when compared

to primed cells [389]. In general, while long-range structural loops don't show great changes, individual contacts between enhancers and cell-type specific genes undergo extensive reorganization when transitioning from naïve to primed cells. This can have widespread effects on the regulation of transcriptional activity and can result in great changes in gene expression patterns during the transition between the two pluripotency states. Indeed, recent studies have shown that there's a very small overlapping percentage of active enhancers between the two pluripotency states, in parallel with extensive changes in the binding profile of essential pluripotency factors such as *OCT4, SOX2* and *NANOG* [345, 390]. Additionally, Polycomb-associated interactions appear to rewire extensively during the naïve-to-primed transition, forming new and more interconnected interaction hubs in primed cells [278].

There is an important need to understand more in depth the dynamics of the interactions between enhancers and their target genes in hPSCs, especially while transitioning between the two pluripotency states. Understanding the global reorganization of enhancer-promoter interactions will represent a step forward in the understanding of the regulation of gene expression programs underlying human development and pluripotency.

1.3.4 The bivalent state of poised enhancers and their role in pluripotency

Poised enhancers (PEs) are a specific class of enhancers that typically display a bivalent signature (**section 1.1.2**). They were first described in 2011, when Rada-Iglesias et al. reported the presence of distal regulatory regions marked by H3K4me1 and H3K27me3 in both mESCs and hESCs [233]. The co-presence of these two marks was later validated by sequential ChIP-seq analysis. Jointly with sub-units of both PRC1 and PRC2, factors associated with active transcription have been found to bind PEs, but the expression levels of PEs associated genes are reportedly low [264]. Initial characterization of PEs revealed that these elements are mainly associated with genes with key roles in processes such as gastrulation, germ layer formation, early somatogenesis and, more generally, associated with development (e.g. *FOX, SOX, WNT*, etc.) and they are a feature that has been observed *in vitro* as well as *in vivo* (mainly in vertebrate genomes) [391].

The term "bivalent domain" (BD) was first introduce when two independent studies, Azuara et al., 2006 and Bernstein et al., 2006, identified loci displaying the co-presence of H3K4me3 and H3K27me3 in mouse ESCs (mESCs) [392, 393]. Bernstein et al, 2006,

profiled H3K4me3 and H3K27me3 patterns of highly conserved non-coding regions by a combination of chromatin immuno-precipitation (ChIP) and DNA array assays. They observed that 75% of TSS regions covered by H3K4me3 were also marked by H3K27me3. The co-presence of the two opposite marks was then confirmed by sequential ChIP experiments which also showed that most of the observed bivalency marked genes encoded for TFs with a major role in development. Despite the presence of the active mark, bivalent genes showed low expression levels in mESCs. Some of them, however, displayed loss of H3K27me3 and higher transcriptional levels upon differentiation, leading to the hypothesis that, perhaps, the role of bivalency lied in the poising of genes for prompt activation at later stages of differentiation [392]. In parallel, Azuara et al., 2006 also provided evidence of the existence of bivalent domains. In their study, they profiled the replication timing of a specific set of genes in mESCs, using replication as a proxy of chromatin accessibility and transcriptional activity. They observed that some genes, despite not being expressed in mESCS, replicated early and were marked by both H3K4me3 and H3K27me3 [393].

Subsequently, several studies profiled bivalent domains genome-wide via ChIP combined with next generation sequencing (ChIP-seq). ChIP-seq profiling revealed that almost all promoters with a high CpG density displayed H3K4me3, 33% of which were also marked by H3K27me3 [392]. BDs were quickly found in hESCs as well and, as seen for mESCS, human bivalent genes were mainly involved in development [394, 395]. It was later shown that BDs were not an artifact resulting from *in vitro* culturing conditions of ESCs, permanently kept in an artificial pluripotency state, and could be identified *in vivo* pluripotent epiblast cells of early post-implantation mouse embryos, as well as in human fetal brain, heart and liver samples [51, 396]. From these studies, bivalency appeared to be rare in mouse oocytes and embryos before implantation, to only emerge at a later developmental stage, in the inner cells mass. After implantation, bivalency was observed in cells of the epiblast and were highly enriched for developmental genes that were previously marked by H3K27me3 in the mouse oocyte and sperm. Thus, bivalency does not seem to be inherited by the zygote, but appears at later stage during early embryogenesis. These findings led to hypothesize that BDs might play a role in committing pluripotent cells to a specific differentiation fate [397, 398, 399, 400].

It has been argued that the observed bivalency could have been a result of heterogeneity of cell population and, since their first report, several studies focused on the characterization of BDs. Experiments on sorted population of T cells, ESCs and embryonic tissues showed that, although there are some cases where heterogeneity can lead to misidenti-

fying bivalent regions, most of them are characterized by the co-existence, indeed, of the two opposite PTMs [401, 394, 402, 403]. More specifically, BDs could exhibit different scenarios where both the active and repressive mark can be present: 1)on the same copy of H3 within the same nucleosome 2)on the same nucleosome, but on different copies of H3 or 3)on two adjacent nucleosomes. Studies combining micrococcal nuclease (MNase) digestion and sequential ChIP for selected genes in C2C12 and mESCs demostrated that, indeed, the two marks co-exist on the same nucleosome [404]. Combining mononucleosome ChIP with mass spectometry (MS), Voigt et al. analyzed bivalency genome-wide and in 2012 they reported that 15% of mononuclesomes carrying H3K4me3 were also marked by H3K27me3, supporting the wide-spread co-existence of the two marks not only for few selected promoters, but for thousands of them. Furthermore, they showed that bivalent domains were characterized by H3K4me3 and H3K27me3 asymmetrically marking opposite copies of H3 within the same nucleosome [405].

TrxG and PcGs are the main players in the establishment of bivalency. In particular, PcG-bound chromatin displays unique properties in ESCs. Developmental genes are predominantly bound by both PRCs and MLL2/KMT2B (both members of TrxG complexes and known "antagonist" of PcGs) which leads to the establishment of BDs (i.e. bivalent promoters and poised enhancers). Work from the Di Croce lab showed that loss of MLL2 in ESCs leads to an increase in occupancy of PcGs at bivalent promoters, with consequent loss of chromatin accessibility and changes in long-range interactions, suggesting that MLL2 and/or its associated PTM, might favor the formation of precise interactions between bivalent promoters in ESCs. Additionally, recent work showed that PcGs are required for the correct activation of genes involved in neural differentiation by providing a permissive chromatin environment at poised enhancers that are already connected to their target genes [264].

CGIs also play an important role in the establishment of bivalency in ESCs. As previously mentioned, most of CGIs at promoters strongly correlate with the presence of H3K4me3 and are devoid of DNA methylation. Indeed, unmethylated CGIs play an important role in the recruitment of TrxG complexes such as MLL1 and MLL2 that are the main responsible for the deposition of H3K4me3 at these specific regions [392, 406, 407]. CpG-rich regions are also involved in determining the composition of histone variants which can also affect bivalency. Indeed, genome wide studies have shown that CGIs determine the presence of variants such as H2A.Z and H3.3 which correlate significantly with

the presence of H3K4me3 at bivalent promoters, as well as enhancers [408, 409, 410, 411]. Likewise, CGIs determine the presence and the maintenance of H3K27me3 at BDs. As observed for H3K4me3, the presence H3K27me3 correlated with CGIs. However, in this case not all CpG-rich regions are marked by H3K27me3, but mainly when they are they associate with lowly transcribed or repressed genes. This led to the hypothesis, further supported by several studies, that as long as CGIs are devoid of transcriptional activators, they contribute to the deposition of both H3K4me3 and H3K27me3, playing a role in the establishment of bivalency. Unmethylated CGIs can, indeed, recruit both MLL and PRC2, mediating the deposition of both marks, and this is usually associated with low transcription or repression [91, 412].

However, the abundance of BDs observed in ESCs leads to hypothesize that other factors, perhaps involved in pluripotency, can play a role in defining the bivalent landscape. As more aspects are being uncovered, more players are likely to be identified. For example, a recent study from the Reik lab identified DPPA2 and DPPA4 as novel factors in the maintenance of both H3K4me3 and H3K27me3 at developmentally relevant promoters [413, 414].

Traditionally, bivalency is considered a hallmark of genes that are set for expression during differentiation. In some cases the bivalent state gets resolved upon differentiation into a monovalent state, keeping H3K4me3 in case of activation or retaining H3K27me3 in case of repression. Ever since the identification of BDs, it has been tempting to speculate that they are restricted to developmental genes and they are involved in an elegant mechanism of regulation whereby genes are held in check until receiving the right cues for their prompt activation or repression. But it is now accepted that bivalency represents a more complex phenomenon, extending to different gene families and found in different cell types [415]. Their functional role is still an area of active research and, if one of the proposed mechanisms is that these specific regions may "poise" genes for rapid activation upon differentiation, additional hypotheses as to why they display this bivalent signature are being explored [416, 417, 418, 415].

Bivalency, however, is not restricted to the co-presence of H3K4me3 and H3K27me3. As briefly mentioned above, over the last decade an additional bivalency signature has been identified that defines **PEs** [233, 243]. Similarly to bivalent promoters, PEs seem to create a chromatin state that can rapidly switch between active and silent. As in the

case of bivalent promoters, it was first hypothesized that PEs resolved their bivalent state upon differentiation to drive the expression of the associated genes in a cell-type specific manner. Indeed, ChIP-seq analysis of differentiated hESCs into neuroectoderm (hNECs) revealed that a good proportion of PEs did display an active enhancer signature in differentiated cells, although a large number of them retained their bivalent nature [233]. In agreement with the change of the PEs state upon differentiation, the genes associated with these elements also showed higher expression levels in hNECs. This was also confirmed by a first GFP (green fluorescent protein) reporter assay where GFP was put under the control of PE elements known to acquire an active state during differentiation. The assay showed that, while GFP expression was undetectable in hESCs, its expression levels increased as hESCS progressed into differentiation to hNECs [233].

These initial findings posited the hypothesis that the bivalent state of PEs gets resolved upon differentiation.

Despite the observation that PEs become active only upon differentiation, 4C-seq and promoter capture-HiC (PCHiC) analyses showed that they establish contacts with their target genes at an earlier stage in ESCs, arguing in favor of their role in the establishment of a permissive chromatin state that allows for readily and robust gene induction [264]. Recent studies have suggested that orphan CGIs, oCGIs,(i.e. intergenic CpG-rich regions that are not associated with annotated promoters) are amongst the genetic elements responsible for the responsiveness of PEs, acting as tethering elements between PEs and their target promoters, also enriched for CGIs. As PEs loop over their target, the transcriptional machinery is delivered to the promoter and it ensures efficient activation. Additionally, it has been suggested that the presence of CGIs at these regions might protect TFs DNA binding motifs from repression by preventing DNA methylation at these regions [265]. The reported data proposed that the combined presence of TFs binding sites and oCGIs at PEs might favor the binding of PcGs. As discussed earlier, PcGs have a known role in the 3D-genome organization and in the establishment of long-range interactions, specifically during development. It is, therefore, plausible to speculate that PcGs might be responsible for keeping in place the contact between PEs and their target, ensuring a prompt activation of their target genes upon differentiation cues.

Furthermore, a recent study from the Rugg-Gunn lab observed that Polycomb- associated interactions hubs show great re-organization between the naïve state of pluripotency and the primed state in human pluripotent stem cells (hPSCs), with the latter showing

an increase in PcGs-bound regions long-range contacts (**Figure 1.10**) [278]. Likewise, mouse pluripotent stem cells (mPSCS) show a reduction of Polycomb-dependent interactions when transitioned to their naïve state.

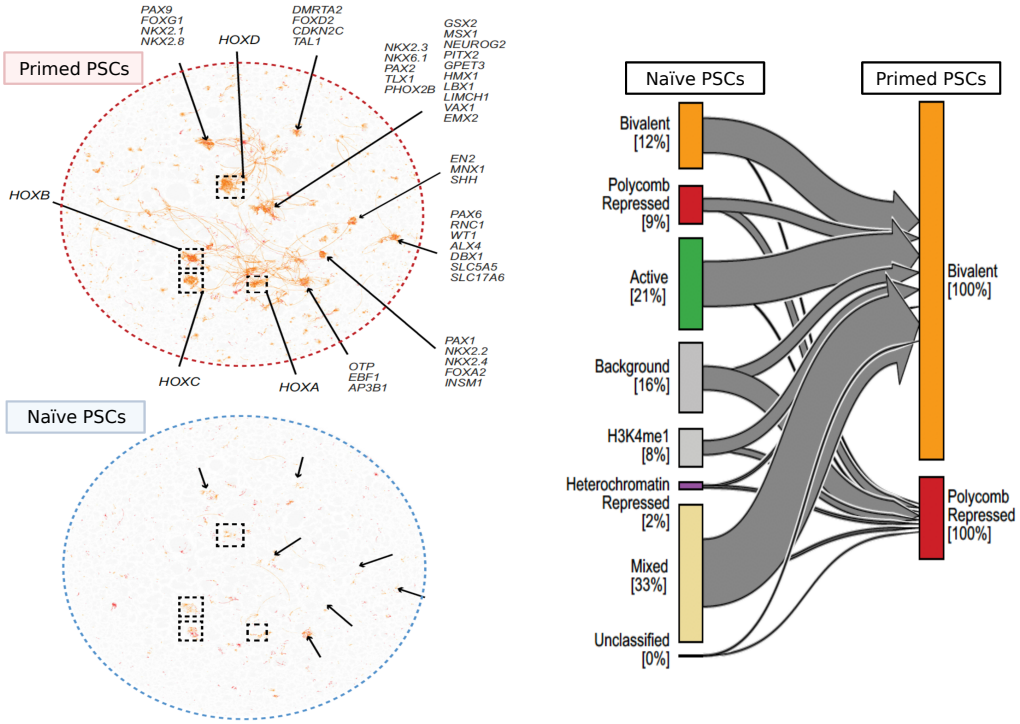


Figure 1.10: **Re-organization of Polycomb-associated interactions hubs between naïve and primed hESCs.** Polycomb-associated interactions undergo great re-organization between the two state of pluripotency, forming numerous interacting clusters in primed hESCs that are not observed in naïve hESCs. Only ∼1/4 of the Polycomb-associated regions in primed hESCs are premarked with H3K27me3 in naïve hESCs, suggesting that the acquisition of the repressive mark occurs as hESCs transition between the two states of pluripotency. Figure adapted from Chovanec,P., et al., 2021 [278].

To uncover the potential functional role of the bivalent signatures represents the next step to gain insight into whether this specific signature represents cause or consequence of their functional role. The continuous advance in gene-editing techniques such as ZFN (Zinc Fingers), TALENS (Transcription activator-like effector nucleases) and, particularly, CRISPR/Cas9-based assays have the potential to untangle bivalency and their downstream functional effects [419, 420, 421, 422].

1.4 THESIS AIMS

Enhancers are key CREs with an essential role in the control of gene expression. Over the last decade a new class of enhancers, **poised enhancers** (PEs), has been identified marked by the joint presence of the active mark H3K4me1 and the repressive H3K27me3. Found primarily in mouse and human ESCs, they engage in DNA looping interactions with key developmental genes and the presence of H3K27me3 in particular, suggests the recruitment of PcGs, which are known to have key role in gene expression control and in mediating long-range interactions in ESCs.

Recent studies suggest that PcGs undergo significant reorganization upon the naïve-to-primed transition, as well as enhancer-promoter contacts, with a high degree of rewiring observed during the transition and can ultimately result in extensive changes of gene expression patterns. Moreover, it has recently been suggested that PcGs may be involved in the establishment and maintenance of contacts between PEs and their target genes in ESCs. Initial studies suggest that while the poised state is prevalent in primed ESCs, it appears less abundant in naïve cells, but it remains unclear if such state is necessary for the priming of human embryonic stem cells and when exactly it is established during the naïve-to-primed transition. Exploring such transition in hESCs opens a window on very early stages of human embryogenesis and development. Given the apparent abundance of PEs in primed hESCs, elucidating the timing of their emergence and their contacts during the transition will represent a step forward in the understanding of the regulation of gene expression programs controlling pluripotency and cell fate.

Through combining experimental and computational analysis, this thesis aims to provide insight into the emergence of PEs during the naïve-to-primed transition and their functional role in the transcriptional control of human pluripotency by:

- devising a chromosome conformation capture approach (PEcHiC) to profile 3D interaction dynamics of PEs;
- profiling PEs chromatin state over time during naive-to-primed transition, in order to understand the relationship between the emergence of the poised state of enhancers and their interaction networks;
- performing CRISPRa-mediated perturbation of candidate PEs in order to gain insight into the functional role of PEs and their role in the regulation of gene expression programs for pluripotency control.

2 Methods

2.1 EXPERIMENTAL METHODS

2.1.1 Human embryonic stem cells (hESCs) culture

WA09 (H9) primed hESCs were provided by WiCell Research Institute, hPSCReg ID: WAe009-A) [366]. Cells have been grown in mTeSR-E8 essential media (StemCell Technologies, Cat.05990), in feeder-free condition using vitronectin XF (xeno-free vitronectin, StemCell Technologies, Cat.07180) as coating matrix. For regular maintenance, cells were passaged every 5 days, at 1:10 ratio, using Gentle Cell Dissociation Reagent (GCDR, StemCell Technologies, Cat.100-0485) and grown at 37°C, 5% CO_2, in normoxia ($O_2 = 20\%$).

Inducible CRISPR-activation (iCRISPRa) WTC11 human iPSCs (hPSCReg ID: UCSFi001-A) were kindly provided by Dr. Martin Kampmann, (University of San Francisco, CA, USA) and generated by the Kampmann's lab as described in Tian R., et al, 2021 [423]. Cells were grown in mTeSR-E8, feeder-free condition, in 6-well plates coated with vitronectin XF. For regular maintenance, cells were usually passaged every 5-6 days, at a 1:10 ratio, using GCDR and grown at 37°C, 5% CO_2, normoxia ($O_2 = 20\%$).

hNES1 hESCs (hPSCReg ID: CAMe001-A) [374] were kindly provided by Dr. Maria Rostovskaya (Babraham Institue, Cambridge, UK). hNES1 naïve cells were cultured and transitioned to the primed state as described in Rostovskaya, M., et al., 2019 [424]. Specifically, the following timepoints were collected over the time course of the transition: **naïve (day 0), day 1, day 3, day 5, day 7, day 10, day 14**. In addition, hNES1 cells were differentiated into definitive endoderm (DE) and neuroectoderm and collected by Dr. M.Rostovskaya (Babraham Institue, Cambridge, UK) as described in Rostovskaya, M., et al., 2019 [424].

2.1.1.1 Freezing and Resuscitating hESCs and iPSCs

H9 hESCs and iCRISPRa WTC11 iPSCs were typically harvested at a confluency of 90-95% using GCDR. Cells were then resuspended in approximately 1x10^6/mL of Freezing

Media (FM): 90% KnockOut Serum Replacement (Gibco, Cat.11520366) + 10% Dimethyl sulfoxide (DMSO, Santa Cruz Biotechnology Cat.67-68-5) + 5μM Y-27632 (ROCKi, Cambridge Bioscience Cat.SM02-1) and were frozen using Corning CoolCell LX Cell Freezing Container (Merk, Cat.CLS432001).

For cell resuscitation, approximately 1x10^6 cells were thawed at 37ºC using a water bath. To dilute the concentration of DMSO, cells were resuspended in 10mL of mTeSR-E8 + 10μM of Y-27632 (ROCKi) and pelleted by centrifuging at RT, 300 x g, for 3 minutes using a swinging bucket centrifuge. After complete removal of DMSO, cells were seeded in mTeSR-E8 + 10μM of Y-27632 (ROCKi) (generally 1x10^6 cells were seeded across 4-5 wells of a 6-well plate) and grown at 37ºC, 5% CO_2, in normoxia (O_2 = 20%). After approximately 24hrs, mTeSR-E8 + Y-27632 (ROCKi) was replaced with fresh mTeSR-E8 media. Cell were then maintained by daily changing media with mTeSR-E8 and passaged as described in the previous section.

2.1.2 Cross-linking of cells

To perform Poised Enhancer Capture Hi-C, hESCs were cross-linked with formaldehyde as follows: cells were washed once with 1X Dulbecco's Phosphate Buffered Saline w/o calcium chloride and magnesium chloride (ModifiedD-PBS, Sigma-Aldrich, Cat.D8537-500ML) and dissociated to a single-cell level with 1X TrypLE Express Enzyme, No Phenol Red (Fisher Scientific, Cat.10718463) at 37ºC for 5 minutes. TrypLE was then neutralized with KnockOut Dulbecco's Modified Eagle Medium (DMEM/F12, Thermo Scientific, Cat.10829018) + 0.1%BSA (Bovin Serum albumin, New England Biolabs, Cat.B9000S). 1x10^6 cells were harvested, resuspended in 5mL 1X D-PBS + 2% Formaldehyde (w/v), Methanol-free (Thermo Fisher Scientific, Cat.28906) and incubated at Room Temperature (RT) for 10 minutes while rotating. To quench formaldehyde and stop fixation, 0.63mL of 1.25M Glycine were added and samples were first incubated for at RT for 5 minutes with rotation followed by an additional incubation of 5 minutes on ice (w/o rotation). Cells were then pelleted for 5 minutes, 500 x g, at 4ºC using a swinging-bucket centrifuge and pellets were resuspended in 5mL of ice-cold 1X D-PBS. After a second step of centrifuging for 5 minutes, 500 x g, at 4ºC, cells were washed with 200uL of ice-cold 1X D-PBS and pelleted for 3 minutes, 8,000 rpm, at 4ºC, using a bench-top centrifuge. After careful removal of supernatant, cells were flash-frozen in liquid nitrogen (N_2) for long-term storage.

2.1.3 Poised Enhancer Capture Hi-C (PECHi-C)

PECHi-C was performed upon a time course of the hNES1 hESCs naïve-to-primed transition (see section **2.1.1** for specific timepoints collected), as well as on hNES1 cells differentiated into definitive endoderm (DE) and neuroectoderm (NE). As a control for the primed state, H9 hESCs were also collected. Two biological replicates for each timepoint of the transition were processed.

LIST of BUFFERS

Hi-C LYSIS BUFFER

10mM	Tris-HCl, pH 8
10mM	$NaCl_2$
0.2%	Igepal CA-630
1X	cOmplete protease inhibitors (EDTA-free)

TWEEN BUFFER (TB)

5mM	Tris-HCl, pH 8
0.5mM	EDTA, pH 8
1M	$NaCl_2$
0.05%	Tween-20

2X No-Tween Buffer, NTB

10mM	Tris-HCl, pH 8
1mM	EDTA, pH 8
2M	$NaCl_2$

Tn5-TAGMENTATION BUFFER

50mM	Tris-HCl, pH 8.4-9
25mM	$MgCl_2$
50%	Dimethyl formamide (DMF)

5X FAST-HYBRIDIZATION BUFFER

1,540mM	$MgC/_2$*$6H_2O$
0.0417%w/w	HPMC
100mM	Tris-HCl, pH 8

WASH BUFFER 1

2X	Tris-HCl, pH 8
0.1%	EDTA, pH 8

WASH BUFFER 2

0.1X	Tris-HCl, pH 8.4-9
0.1%	$MgCl_2$

2.1.3.1 Hi-C stage

Hi-C represents the first step of the Capture Hi-C protocol. Formaldehyde-fixed cells (see **section 2.1.2**) for each timepoint of the transition were resuspended in 100μl of Hi-C

lysis-buffer and incubated on ice for 30 minutes for gentle cell permeabilization. Using a swinging-bucket centrifuge, cells were pelleted at 4°C, 300 x g, for 5 minutes. Following two washes with 1.2X NEBuffer 3 (New England Biolabs, Cat.B7003S), cells were resuspended in a final volume of 350μl of NEBuffer 3 + 12μl of 10%SDS and incubated at 37°C for 1 hour. 80μl of 10%Triton X-100 were then added to the samples which were incubated at 37°C for an additional hour. Restriction digestion of chromatin was carried out by adding 100U of *DpnII* enzyme (New England Biolabs, Cat.R0543T) and incubation at 37°C, 950rpm, for 16 hours.

Following chromatin digestion with *DpnII*, the resulting restriction fragments were labelled with biotin by adding:

4.5μl	10mM dCTP/dGTC/dTTP mix (Thermo Fisher, Cat.10297117)
15μl	1mM biotin-14-dATP (Jena Bioscience, Cat.NU-835-BIO14-S)
10μl	DNA Polymerase I, Large (Klenow) Fragment (5 U/uL) (New England Biolabs, Cat.M0210L)
22μl	TLE buffer (10mM Tris-HCl, pH 8, 0.1mM EDTA, pH 8)

and incubated at 37°C for 45 minutes with intermittent shaking: 700rpm for 10 seconds, every 30 seconds. Biotinylated digested chromatin was then pelleted at 4°C, 600 x g, for 6 minutes and ligation reaction was performed by the addition of:

100μl	10X T4 DNA ligase buffer (New England Biolabs, Cat.B0202S)
5μl	BSA, 20mg/mL (New England Biolabs, cat.B9000S)
10μl	T4 DNA ligase (Invitrogen, Cat.15224017)
835μl	H_2O

and incubation at 16°C for 4 hours. Samples were centrifuged at 4°C, 600 x g, for 6 minutes and after removal of 800μl of ligation reaction, chromatin was resuspended in the remaining 200μl. De-crosslinking was performed at 65°C, o/n + 15μl of 10mg/ml Proteinase K (Roche, Cat.03115879001). After initial 2-hour incubation, an additional 15μl of 10mg/ml Proteinase K were added and the de-crossilinking reaction was incubated at 65°C, o/n.

Following de-crosslinking of the chromatin, Hi-C DNA was purified by performing 1X AmpureXP (Beckman Coulter, Agencourt AMPureXP Cat.A63880) DNA clean up. In

order to allow the retrieval of very long fragments typical of the Hi-C material at this stage, ligated Hi-C DNA was incubated with AmpureXP beads for 10 minutes at RT. Following two washes with 70% Ethanol (EtOH), AmpureXP were air-dried and Hi-C chromatin was eluted in a final volume of 30μl of DNAse/RNase-free H_2O, for 10 minutes. Hi-C DNA was then quantified using Qubit DNA HS assay (Thermo Fisher Scientific, Cat.10616763) as per manufacturer's instructions.

For Hi-C library amplification, Hi-C ligated DNA was tagmented using home-made Tn5 enzyme pre-loaded with Illumina sequencing adaptors (see **Appendix C**) or Illumina Tagment DNA TDE1 enzyme (Illumina, Cat.20034198). Tagmentation reaction was set up on ice using 100ng of Hi-C DNA per reaction and an average of 8 tagmentation reactions per sample were set up. After Tn5-mediated tagmentation at 55°C for 7 minutes, 1/4 volume of 0.2%SDS was added and reaction was incubated for an additional 7 minutes at 55°C. Prior to pulldown, 25μl of Streptavidin MyOne C1 Dynabeads (Invitrogen, Cat.65001) per sample were washed twice with TB Buffer. For each wash in TB, beads were incubated on tube rotator for 3 minutes at RT. Following TB washes, beads were resuspended in 25μl of 2X NTB buffer. After incubation at 55°C, biotin-streptavidin pulldown was carried out at RT for 45 minutes, by adding tagmented Hi-C DNA to prewashed Streptavidin MyOne C1 Dynabeads. To remove non-biotinylated DNA carryover, beads were then washed with with 1X NTB (1:2 dilution of 2X NTB Buffer) at 65°C for a total of three washes (for each wash beads in 1X NTB were incubated at 65°C for 3 minutes). Following 1X NTB washes, two additional washes in 1X TLE were performed at 65°C, 1 minute incubation for each wash. Hi-C DNA bound to streptavidin beads was then resuspended in 25μl of DNAse/RNAse-free H_2O and used for Hi-C library Polymerase Chian Reaction (PCR) amplification using KAPA HiFi plus dNTPs kit (Roche, Cat.7958838001). Each PCR reaction was set up as follows:

10μl	5x KAPA HiFi buffer
1.5μl	10mM dNTPs
1.5μl	10μM i7 primer
1.5μl	10μM i5 primer
1μl	KAPA HiFi DNA polymerase

and for each Hi-C samples, a total of five PCR reactions were set up using 5μl of beads bound to Hi-C DNA and Hi-C library was amplified following the thermocycling settings

below:

Temperature	Time	n amplification cycles
72°C	3 minutes	1
95°C	30 seconds	1
95°C	10 seconds	5
55°C	30 seconds	
72°C	1 minute	
72°C	1 minute	1

PCR reactions for each sample were then combined and DNA was purified by performing a 1X AmpureXP DNA clean up. AmpureXP beads + DNA were incubated for 10 minutes at RT and, after two washes with 70% EtOH, AmpureXP beads were air-dried and Hi-C library was eluted in a final volume of 30μl of DNAse/RNase-free H_2O, for 10 minutes. The final Hi-C library was then quantified using Qubit DNA HS assay (Thermo Fisher Scientific, Cat.10616763) as per manufacturer's instructions and D1000-4200 Agilent Tapestation (Agilent, Cat.G2991BA), following manufacturer's instructions (**Figure 2.1**).

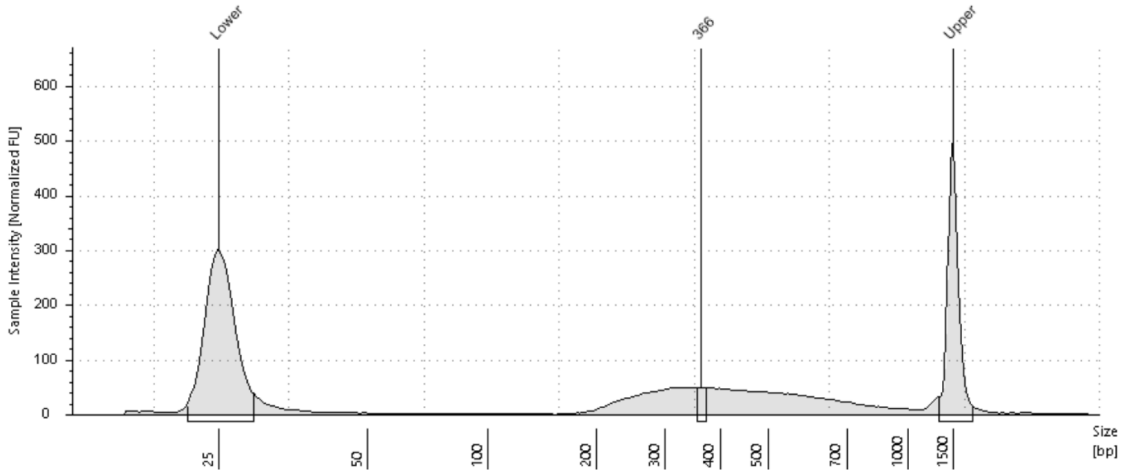


Figure 2.1: **Example of D1000-4200 Agilent Tapestation profile of a typical Hi-C library.** A typical D1000-4200 Agilent Tapestation profile of a Hi-C library after PCR amplification, showing a fragment size distribution between 300bp and ≥1,000bp. Hi-C library generated starting from 1x10^6 primed hESCs (H9).

2.1.3.2 Capture Hi-C stage

To perform hybridization of Hi-C libraries with RNA-biotinylated probes, early timepoints (i.e. naïve, day 0, day 3, day 5, day 7, day 10) and late timepoints (day 14, primed H9 hESCs, DE and NE) for each biological replicate were pooled together so that each capture reaction received between 250ng and 450ng of total Hi-C DNA template as starting material. Pools were balanced based on quantitative PCR (qPCR) based quantification of Hi-C libraries using the KAPA Library Quantification kit (Roche, Cat.07960140001) for NGS ready libraries so that each sample was represented equally within the pool. Hi-C DNA libraries were concentrated and resuspended in 11μl of DNAse /RNAse-free H_2O using Eppendorf Concentrator plus (Eppendorf, Cat.SKU - 12305985305 000568). A blocker mix containing human Cot-1 DNA (Agilent, Cat.5190-3392) + sonicated salmon sperm DNA (Agilent, Cat.201190-81) + 2,350μM custom blockers (**Appendix C**) was added to the concentrated Hi-C DNA library and the samples were incubated using the following thermocycling conditions:

Temperature	Time
95ºC	5 minutes
65ºC	10 minutes

The hybridization mix was then added to the samples, as described below:

10μl 1:4 diluted SureSelect RNase block (Agilent, cat.300151)

1.5μl SureSelect Capture system

1.5μl 5X fast hybridisation buffer

and incubated in the following thermocycling conditions:

Temperature	Time	n amplification cycles
65ºC	1 minute	1
65ºC 65ºC	1 minute 3 seconds	60
65ºC	1 minute	1

During hybridization, 60μl of Dynabeads MyOne Streptavidin T1 beads (Invitrogen, Cat.65601) per sample were washed three times with 200μl SureSelect Binding Buffer

(Agilent, SureSelect XT HS2 DNA system), incubating samples for 5 minutes at RT for each wash. Following the third wash, Dynabeads MyOne Streptavidin T1 beads were resuspended in 200μl of SureSelect Binding Buffer per sample. After completion of the hybridization step, samples were added to pre-washed Dynabeads MyOne Streptavidin T1 beads to perform biotin-streptavidin pulldown at RT for 30 minutes. To remove non-specifically bound probes, pulldown reactions were then washed twice with Wash Buffer 1 at RT, 10 minutes incubation per wash. RT washes were then followed by washes with Wash Buffer 2 pre-warmed at 68°C, for a total of three washes. For each wash in Wash Buffer 2, samples were incubated at 68°C for 10 minutes with intermittent shaking at 700rpm for 5 seconds, every 2.5 minutes. Following washes with Wash Buffer 1/2, beads bound to the Capture Hi-C library were resuspended in 25μl of DNAse/RNAse-free H_2O and used for PCR library amplification using KAPA HiFi plus dNTPs kit. As for Hi-C library PCR amplification, for each sample five PCR reactions using 5μl of beads + DNA were set up as follows, in a total final volume of 50μl:

10μl	5x KAPA HiFi buffer
1.5μl	10 mM dNTPs
1μl	10μM FCA-P7F
1μl	10μM P5-FCA-R
1μl	KAPA HiFi DNA polymerase

and amplified following the thermocycling conditions below:

Temperature	Time	n amplification cycles
95°C	3 minute	1
95°C	20 seconds	5
55°C	30 seconds	
72°C	30 seconds	

PCR reactions for each samples were then pooled together and the supernatant containing the amplified Capture Hi-C library was separated from the beads and retrieved. The Capture Hi-C library was then purified performing a 1X AmpureXP DNA clean up. Following incubation at RT for 10 minutes, two washes with 70% EtOH were performed and AmpureXP were air-dried. Capture Hi-C DNA library was eluted in a final volume of 25μl of DNAse/RNase-free H_2O, for 10 minutes. Final libraries were then quantified

with qPCR quantification using the qPCR-based KAPA Library Quantification kit for NGS ready libraries (for each Capture Hi-C library a 1:1000 and 1:2000 was used for qPCR-based quantification) and D1000-4200 Agilent Tapestation (Agilent, Cat.G2991BA), following manufacturer's instructions (**Figure 2.2**).

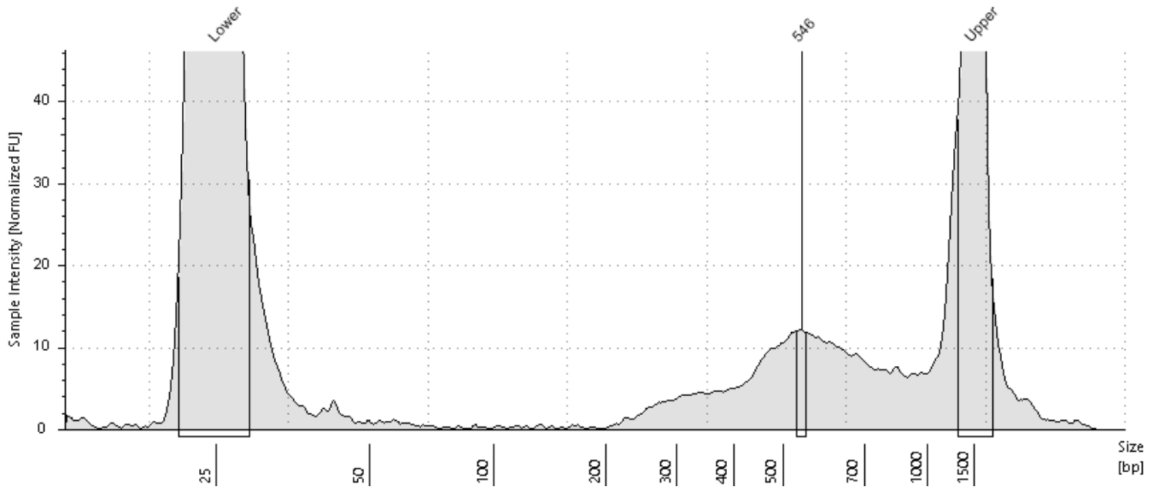


Figure 2.2: **Example of D1000-4200 Agilent Tapestation profile of a typical CHi-C library.** A typical D1000-4200 Agilent Tapestation profile of a CHi-C library after PCR amplification, showing a fragment size distribution between 300bp and ≥1,000bp. Often CHi-C libraries show a shift, or "bump", towards larger fragments, relative to the starting Hi-C library. CHi-C library generated starting from ∼500ng of Hi-C material generated from 1x10^6 primed hESCs (H9).

Final Capture Hi-C libraries were then multiplexed in an equimolar pool at a concentration of 1nM and sequenced at Novogene Ltd (Cambridge, UK) on a NovaSeq6000 Illumina sequencing platform, flowcell S4, 150bp paired end (PE), for a total of ≥ 2 billion reads.

2.1.4 Cleavage Under Targets and Tagmentation (Cut&Tag) technology chromatin profiling

Cut&Tag (C&T) is a newly developed method that allows to profile DNA-binding proteins starting from unfixed, live permeabilized cells [425]. Briefly, after permeabilization, cells are immobilized on Concanavalin A (ConA) magnetic beads and are incubated with an antibody (Ab) that recognizes a protein or a histone modification of interest (primary Ab). Cells are then incubated with a secondary Ab that recognize the heavy chain of the primary Ab, followed by incubation with Tn5 transposase enzyme conjugated to NGS adaptors fused to protein A/G which carries out Ab-targeted tagmentation. Cleaved DNA

is then extracted and used for DNA-library amplification followed by high-throughput sequencing.

C&T was performed upon a time course of the hNES1 hESCs naïve-to-primed transition (see section **2.1.1** for specific timepoints collected), as well as on hNES1 cells differentiated into definitive endoderm (DE) and neuroectoderm (NE). As a control for the primed state, H9 hESCs were also collected. Three biological replicates for each timepoint of the transition were processed.

LIST of BUFFERS

BINDING BUFFER

20mM	HEPES pH 7.5
10mM	KCl
10mM	$CaCl_2$
10mM	$MnCl_2$

WASH BUFFER

20mM	HEPES pH 7.5
150mM	$NaCl_2$
0.5mM	spermidine
1x tablet	cOmplete protease inhibitors (EDTA-free)

ANTIBODY BUFFER

2mM	EDTA pH8
0.1%	BSA
up to volume	DIG-WASH BUFFER

DIG-WASH BUFFER

0.05%	digitonin
up to volume	WASH BUFFER

TAGMENTATION BUFFER

10mM	$MgCl_2$
up to volume	DIG-300 WASH BUFFER

DIG-300 WASH BUFFER

20mM	HEPES pH 7.5
300mM	$NaCl_2$
0.5mM	spermidine
1x tablet	cOmplete protease inhibitors (EDTA-free)
0.01%	digitonin

2.1.4.1 Immuno-precipitation and DNA tagmentation

C&T chromatin profiling assay was performed on unfixed, live hNES1 and H9 hESCs. Firstly, 10μl of ConA beads (Stratech, Cat.BP531-BAN-3ml) per 100K cells were activated by washing the bead slurry twice in 800uL of Binding Buffer in DNA LoBind 1.5mL tubes (Scientific Laboratory Supplies, Cat.E0030108051). ConA beads were then resuspended in 10μl of Binding Buffer per 100K cells and they were kept on ice while harvesting cells.

After two washes with 1X D-PBS (Sigma-Aldrich, Cat.D8537-500ML), cells were dissociated at a single cell levels with 1X TrypLE Express Enzyme, No Phenol Red at 37°C for 5 minutes. TrypLE was then neutralized with KnockOut DMEM/F12 + 0.1%BSA. After removal of TrypLE, cells were washed once and resuspended in 1ml of Wash Buffer. Cells were then added, drop-wise, to previously activated ConA beads while gently vortexing (1100rpm). ConA beads + cells were incubated on tube rotator, at RT for 10 minutes. Following the binding of the cells, ConA beads were resuspended in ice-cold Antibody Buffer, 100μl for Immuno-Precipitation (IP) reaction, transferred in 1.5mL tube (not LoBind) and antibodies (Ab) for the targets of interest were added as described in **Table 2.1**.

Target	Concentration	Source
H3K4me1	1:100	Abcam, Cat.ab8895
H3K4me3	1:100	Active Motif, Cat.39159
H3K27me3	1:100	Cell Signaling Technology, Cat.9733
H3K27ac	1:100	Millipore, Cat.MABE647
p300	1:100	Cell Signaling Technology, Cat.D2X6N
BRD4	1:100	Active Motif, Cat.91302
IgG (negative control)	1:100	Abcam, Cat.ab2410

Table 2.1: **Antibodies used for the C&T immuno-precipitation experiments.** The table summarizes the specification of the antibodies used for the C&T immuno-precipitation reaction and the concentration used.

Following o/n IP at +4°C, samples were washed once in Dig-Wash Buffer and re-suspended in a 100μl of 1:100 secondary Ab (guinea pig α-rabbit IgG, Rockland, Cat. ABIN101961) + Dig-Wash Buffer and incubated on tube rotator, at RT for 30 minutes. After three washes in Dig-Wash Buffer, a 1:250 dilution og the pAG-Tn5 construct (CUTANA pAG-Tn5 for CUT&Tag, EpiCypher, Cat.15-1117) + Dig300-Wash Buffer was then added and samples were incubated at RT for 1 hour to allow for the binding of pAG-Tn5 to the secondary Ab at the regions of interest. After incubation with pAG-Tn5, three washes in Dig300-Wash Buffer were carried out and the tagmentation reaction was activated by re-suspending the samples in Tagmentation Buffer and incubating at 37°C for 1 hour. The tagmentation reaction was then stopped by the addition of 10μl 0.5M EDTA, 3μl 10% SDS and 2.5μl 20mg/ml Proteinase K and incubation at 55°C for 1 hour. The immuno-precipitated and tagmented DNA was purified by performing a 1X AmpureXP DNA clean up to isolate DNA fragments with size $\geq$100bp and extracted DNA was eluted in a final volume of 45μL of deionized water (diH_2O).

2.1.4.2 Library amplification and sequencing

C&T libraries were amplified using the KAPA HiFi DNA Polymerase HotStart kit (Roche Sequencing Store, Cat.KK2601). Briefly, 21μl of purified DNA was added to 25μl of KAPA HiFi HotStart Ready Mix and i7/i5 primer mix containing iNext unique 8bp barcodes for library multiplexing in a 50μl final reaction. After activation of the KAPA HiFi DNA Polymerase enzyme at 95°C, for 3 minutes, C&T libraries were amplified using the following thermocycler conditions:

Temperature	Time	n amplification cycles
72°C	5 minutes	1
98°C	30 seconds	1
98°C	10 seconds	n
60°C	10 seconds	
72°C	1 minutes	1

Specifically, DNA isolated from H3K4me1, H3K4me3 and H3K27me3 IP reactions received n = 11 amplification cycles, while DNA from H3K27ac IP reactions received n = 14 amplification cycles. DNA isolated from p300 and BRD4 IP reactions received a total of n = 18 amplification cycles, as well as DNA isolated from the negative control α-IgG IP reaction. The number of amplification cycles was empirically determined based on profiles and concentration of C&T libraries resulting from an initial amplification of 13 cycles. A total of 213 C&T libraries were processed and sequenced. Final libraries were quantified using Agilent 2100 Bioanalyzer profiles or D1000-4200 Agilent Tapestation (**Figure 2.1**) and qPCR-based quantification using the KAPA Library Quantification kit for NGS ready libraries.

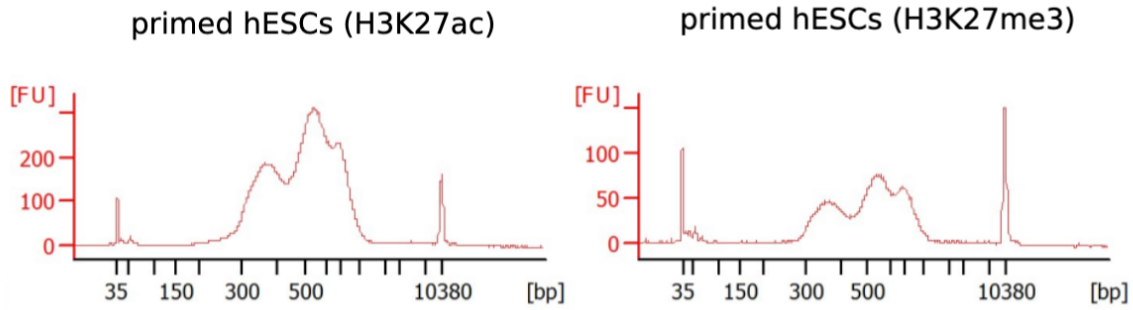


Figure 2.3: **Example of a Cut&Tag Agilent 2100 Bioanalyzer library profile.** Agilent 2100 Bioanalyzer profiles of H3K27ac and H3K27me3 C&T libraries in primed hESCs. The peak at $\sim$ 350bp represents mono-nucleosomes and larger fragment size peaks represent the presence of oligo-nucleosomes in the library.

Final libraries were then multiplexed in an equimolar pool at a concentration of 1.5nM and sequenced at Novogene Ltd (Cambridge, UK) on a NovaSeq6000 Illumina sequencing platform, flowcell S4, 150bp paired end (PE), for a total of $\geq$6 million reads for each library.

2.1.5 Inducible CRISPR-activation (iCRISPRa)

2.1.5.1 sgRNAs designing and cloning

Plasmids for the expression of sgRNAs for *NEUROD1* promoter and scramble sgRNAs were kindly provided by Dr. Andrew Bassett (Sanger Institute, Cambridge, UK), while previously published sgRNAs were used to target *CXCR4* promoter and *GATA1* promoter and enhancers [423, 426] (for a complete list of sequences see **Appendix A**).

sgRNAs were individually cloned in pGL3-U6-sgRNA-PGK-puromycin plasmid, a gift from Xingxu Huang (Addgene plasmid #51133; http://n2t.net/addgene: 51133; RRID: Addgene _51133). The plasmid backbone contains a sgRNA scaffold and a 20bp stuffer flanked by two *BsaI* restriction sites, under the control of the U6 promoter. In addition, it contains and ampicillin resistance gene, for selection of successful cloning of sgRNAs, and a puromycin resistance gene, for selection of successfully transfected hESCs.

Briefly, 1.5μg of pGL3-U6-sgRNA-PGK-puromycin plasmid was digested of with *BsaI* enzyme (New England Biolabs, Cat.R0535S) at 37°C, 50 minutes. To de-phosphorylate the "sticky-ends" resulting after digestion with *BsaI*, 1uL of Quick CIP (New England Biolabs, Cat.M0525S) was added to the reaction for further incubation at 37°C for 10 minutes (total time of digestion: 1 hour). To purify the linear plasmid and remove the excised 20bp stuffer (hence, further minimizing plasmid's re-ligation events), the digested plasmid was purified using the QIAquick PCR Purification Kit (QIAGEN, Cat.Cat No./ID: 28104) following manufacturer's protocol.

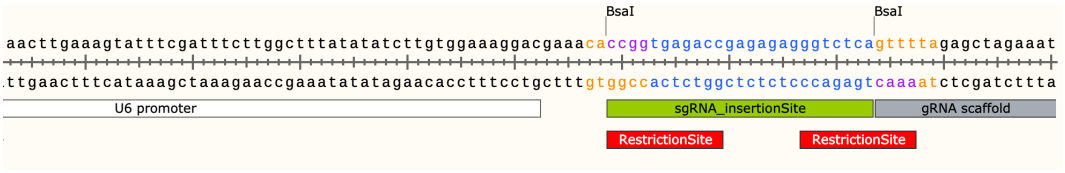


Figure 2.4: **Schematic of pGL3-U6-sgRNA-PGK-puromycin sgRNA cloning site.** *BsaI* plasmid's over-hangs represented in orange. In purple the "sticky-ends" added to the sgRNA oligos are complementary to the plasmid's over-hangs following the *BsaI* restriction digestion (orange). In blue, the site of the specific 20bp sgRNA sequences.

All sgRNA oligos were designed bearing complementary overhangs to the plasmid's "sticky-ends" generated after digestion with *BsaI* (**Figure 2.4**) and 100μM of each sgRNA forward (F) and reverse (R) oligos were annealed using the thermocycling conditions below:

Temperature	Time
37°C	30 minutes
95°C	5 minutes
95°C	5 minutes
ramp down 5°C/min	
25°C	hold

50ng of linearized plasmid was then ligated at 16°C, o/n with the annealed sgRNA oligos (1:10 dilution) using 1μl of 10,000 Units/mL T4 ligase (New England Biolabs, Cat. M0201S). The ligated product was then used for transformation of *E.coli*, Stbl3 competent cells.

Briefly, Stbl3 competent cells were made starting from 50uL of Stbl3 competent cells sampler (Thermo Fisher Scientific, Cat.A10469). A single colony of Stbl3 cells grown on LB (Luria Broth, provided by the MRC-LMS) was inoculated in 25mL of LB broth and grown at 37°C, 180rpm (rotations per minute), o/n. The 25mL starter culture was then diluted to a starting OD_{600}=0.20 and incubated at 37°C, 180rpm until culture reached OD_{600}=0.4 (i.e. exponential or logarithmic phase of growth). Cells were then chilled on ice for 15 minutes, pelleted at 4°C, 1000 x g for 10 minutes and resuspended in ice-cold 0.1M $CaCl_2$ and incubated at 4°C for 30 minutes. $CaCl_2$ is known to increase the efficiency of the uptake of DNA during transformation of bacteria cells. The divalent $Ca_2{}^+$ cations generated transiently create pores on the bacteria cell wall which facilitate the entry of foreign DNA in the cell. After incubation, cells were pelleted at 4°C, 1000 x g, 10 minutes and resuspended in ice-cold 0.1M $CaCl_2$ + 15% glycerol. 50μl aliquots of competent Stbl3 cells were then used for transformation with ligated plasmid containing sgRNAs.

For Stbl3 cells transformation, 2μl of ligated product was added and cells were incubated for 20 minutes on ice. Cell underwent heat shock at 42°C for 30 seconds and incubated on ice for an additional 2 minutes, which enables the uptake of DNA by the cells in a calcium-rich environment by counteracting the electrostatic repulsion between the plasmid DNA and the bacterial cellular membrane. Transformed cells were then grown in 1ml of SOC outgrowth medium (Super Optimal broth with Catabolite repression, New England Biolabs, Cat.B9020S) at 37°C for 1 hour. 950μl of SOC media was when removed and the remaining 50μl was plated on LB agar plates containing 100μg/ml ampicillin. Plates were incubated at 37°C o/n to allow successfully transformed cells carrying the ampicillin resistance gene to grow.

For plasmid isolation, single colonies (typically, 2 colonies for each sgRNA were picked) were then used to inoculate 5ml LB + 100μg/ml ampicillin starter cultures and incubated at 37°C, 180rpm, o/n. 4ml of the o/n starter cultures were then used for plasmid isolation following the QIAprep Spin Miniprep Kit (QIAGEN, Cat.27106) or the EZNA KIT ENDO-FREE PLASMID MINI I (VWR, Cat.D6948-01) following manufacturer's instructions. The correct ligation of sgRNAs into the plasmid was then checked by Sanger Sequencing (outsourced to GeneWiz, Azenta Life Sciences). Briefly, between 50ng/μl and

80ng/μl of isolated plasmid was mixed with 5μM of the hU6-F primer in a final reaction volume of 10μl at a 1:1 ratio (for complete primer sequence see **Appendix C**).

2.1.5.2 iCRISPRa transfection of iPSCs

WTC11 iPSCs were transfected using TransIT-LT1 transfection reagent (Mirus Bio, Cat. MIR2300). A total of 4μg of plasmid DNA was incubated with 12μl of TransIT-LT1 transfection reagent at RT for 20 minutes followed by incubation at 37°C for an additional 20 minutes. 180K cells for each transfection reaction were dissociated at a single cell level using Accutase cell dissociation reagent, 1mL/well (Fisher Scientific, Cat.11599686) at 37°C for 5 minutes. Cells were then harvested, seeded in E8 + 10μM ROCKi containing transfection complexes and incubated at 37°C, 5%CO_2, normoxia levels ($O_2 = 20\%$). Cells were then treated with 0.5μg/mL puromycin (2mL of mTeSR-E8 + 0.5μg/mL puromycin per well) to select for successfully transfected cells for 48hrs, followed by 24hrs recovery in mTeSR-E8 media. In order to counteract the degradation of DHF-dCas9-VPH mediated by dehydrofolate reductase enzyme (DHFR), mTeSR-E8 media was replaced with mTeSR-E8 + 20μM trimethoprim (TPM). After 24hrs treatment with TMP, cells were harvested for RNA extraction.

In each experimental settings cells were transfected with sgRNAs specific to target genes of interest (i.e. *NEUROD1*, *CXCR4*, *GATA1*) and the following three controls: 1) untransfected cells not treated with TMP (-TMP); 2) untransfected cells treated for 24hrs with TMP (+TMP); 3) cells with TransIT-LT1 transfection reagent only (w/o DNA) and treated with TMP (Mock).

2.1.6 RNA extraction

After two washes in 1X D-PBS, cells were resuspended in 400μl of RLT buffer (from RNeasy Mini Kit, QIAGEN, Cat.74106) and lysed using QIAshredder biopolymer shredding system (QIAGEN, Cat.79656) following manufacturer's instructions. RNA was extracted from lysate using RNeasy Mini Kit, following manufacturer's protocol. For complete DNA removal, QIAGEN RNase-Free DNase Set (QIAGEN, Cat.79254) was used to perform DNA digestion on column, as per manufacturer's instructions. Final RNA was quantified using NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Scientific), measuring sample's absorbance at 260nm: yield for all samples ranged between 700ng and 1μg, with 260nm/280nm ratio $\geq$1.9, indicative of good quality RNA.

2.1.7 RNA Retro-transcription

Between 300ng and 700ng of extracted RNA was used for retro-transcription and generation of single-stranded cDNA for following quantitative real-time polymerase chain reaction (qRT-PCR) assay. Retro-trascription was carried out using superscript IV Reverse Transcriptase enzyme (Fisher Scientific, Cat.18-090-200) as per manufacturer's instructions. Final cDNA was quantified using NanoDrop One Microvolume UV-Vis Spectrophotometer measuring sample's absorbance at 260nm: cDNA final yield ranged between 700ng and 900ng, with 260nm/280nm ration $\geq$1.8, indicating a good level of cDNA purity.

2.1.8 Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

To asses gene activation in the iCRISPRa WTC11 iPSCs, single stranded cDNA was used to perform qRT-PCR using Applied Biosystems PowerUp SYBR Green Master Mix (fisher Scientific, Cat.15350929). Three technical replicates for each reaction were set up along with three replicates of No Template Control (NTC) to identify reaction contaminants for each primer pair. qRT-PCR reactions were set up in a 10μl final reaction volume,as follows:

PowerUp SYBR Green Master Mix (2X)	5μl
Forward and Reverse Primers (400nm each)	0.8μl
cDNA template (7ng, $\leq$10% of total volume)	2μl
H_2O	2.2μl

and qRT-PCR amplification was carried out in the thermalcycling conditions below, using the QuantStudio™ Real-Time PCR System (Thermo Fisher Scientific):

Temperature	Time	Cycles
50°C	2 minutes	1
95°C	2 minutes	1
95°C	3 seconds	40
60°C	30 seconds	

$\Delta\Delta C_t$ method was applied to calculate fold change of gene expression. Briefly, the average C_t for the housekeeping gene *GAPDH* (H) and for the target genes (T) were

calculated for controls (C) and experimental (E) samples. Gene expression fold change was then calculated as follows:

$$(\mathbf{1})\ \Delta C_t(C) = TC - HC$$

$$(\mathbf{2})\ \Delta C_t(E) = TE - HE$$

$$(\mathbf{3})\ \Delta\Delta C_t = \Delta C_t(E) - \Delta C_t(C)$$

$$(\mathbf{4})\ 2^{-\Delta\Delta C_t}$$

The final value $2^{-\Delta\Delta C_t}$ represents the fold change of the expression of the targeted genes after dCas9-VPH activation relative to the controls, after normalization to the housekeeping gene. For a complete list of the primers used in the qRT-PCR assay see **Appendix C**.

2.1.9 RNA Flow-FISH assay

RNA Flow-FISH assay was performed using PrimeFlow^{TM} RNA Assay Kit (Invitrogen, Cat.88-18005-204). The PrimeFlow assay is an in situ hybridization assay that combines the branched DNA-technology with the single cell resolution of flow cytometry. Particularly, branched-DNA technology matches target-specific probes to amplify the detection of a specific RNA transcript, achieving between 8,000 to 16,000 fold signal amplification.

The assay was performed in hESCs following the manufacturer's protocol, changing all centrifuging steps which were performed at 300 x g instead on 800 x g. Briefly, after fixation and permeabilization in suspension, 1x10^6 cells per target gene were incubated with the following gene-specific target probes:

Target	Type	Specification	Cat.N^o
GAPDH	Type 1	Alexa Fluor 647	PF-204
OCT4 (*POU5F1*)	Type 1	Alexa Fluor 647	PF-210
NEUROD1	Type 6	Alexa Flour 750	PF-204
LHX6	Type 4	Alexa Fluor 488	PF-204

After hybridization with gene-specific probes, in order to achieve amplification of the signal, cells were then incubated with pre-amplifier and amplifier probes that will function as a scaffold for the hybridization of fluorescent label probes. Cells were then process

through flow cytometry to detect the cell population expressing the specific genes of interest.

Analysis of flow-cytometry data we performed in FlowJo, by the Babraham Institute FlowCytometry facility. Flow cytometry was performed using a BD LSRII analyzer (BD Bioscience) using the following parameters:

Target	Laser line	Emission
GAPDH	640nm (red laser)	660/20
OCT4 (*POU5F1*)	640nm (red laser)	660/20
NEUROD1	633nm (red laser)	780/60
LHX6	488nm (red laser)	525/25

2.2 COMPUTATIONAL ANALYSES

All the analyses were carried out in the Shell environment or R statistical environment, unless stated otherwise. (Code availability: https://github.com/MoniDR/PECHiC).

2.2.1 Identification of Poised Enhancers (PEs) regions and design of PE capture system

PEs were identified based on their chromatin profile and the joint presence of H3K4me1 and H3K27me3 signals. For this purpose, six published ChIP-seq datasets were used (ref: The ENCODE Project Consortium; Vallot C., et al., 2015; Roadmap Epigenomics Project; Rada-Iglesias A. et al., 2011) in combination with unpublished Cut&Run (C&R) data for H3K4me1 and H3K27me3 in H9 primed hESCs (Rugg-Gunn lab, Cambridge). Specifically, bam files were downloaded and aligned with bowtie2 to the GRCh38 reference genome and using the following parameters: `bowtie2 -p 8 --no -unal -t --phred33 -quals` [427]. Peaks were called using Macs2 peak-caller applying an initial lenient cut-off of p-value = 10^{-2} as follows: `macs2 callpeak -g hs -p 0.01 --nomodel --shift 0 --extsize 200/300 -B --SPMR --keep-dup all [--broad]` (`broad` option was used to call H3K27me3 specifically) [428]. *ChIPQC* Bioconductor R package [429], v.1.21.0, was used to compute quality metrics of the ChIPseq datasets and those with a FRiP (Fraction of Reads in Peak) $\leq$10% were excluded from the analyses. Peak cutoff was later adjusted for ChIP-seq and C&R datasets based on precision-recall analysis. Specifically, PE regions

identified based on the combinations of ChIP-seq and C&R data were compared to PE regions identified in the original publication [233], either through direct overlap of ChIP-seq peaks or ChromHMM integration of the two histone marks at promoter interacting regions (PIRs) defined in [430, 277]. Based on the precision-recall analysis, more stringent cutoffs were then applied to ChIP-seq and C&R dataset: p-value=$10^{-6.5}$ and p-value=10^{-5}, respectively, and a combination of bivalent regions identified by both ChIP-seq and C&R were included into the design.

In order to design the Poised Enhancer Capture Hi-C capture system, once PE regions were identified, they were then mapped to the *DpnII*-digested genome fragments in order to design complementary biotinylated RNA probes using a customized script (provided by Babraham Bioinformatics, Babraham Institute, Cambridge UK). RNA probes were then generated and provided by Agilent Technologies (SureSelect Custom DNA Target Enrichment Probes, 6.0 - 11.999 Mbp).

2.2.2 Poised Enhancer Capture Hi-C (PECHi-C) data processing

2.2.2.1 Data alignment and pre-processing

PECHi-C sequencing data were aligned and pre-processed using the HiCUP pipeline [431]. Specifically HiCUP truncates paired-end sequencing reads from the 5′ end at a ligation junction, creating single-end reads that are then individually aligned to the genome through bowtie/bowtie2 [427]. The individual reads are then re-paired or discarded if one of the reads of the pair did not map correctly to the genome and HiCUP then applies various filters to discard invalid di-tags that could results from scenarios such as same-fragment ligation, adjacent ligation and/or re-ligation events. At the end of the alignment process, HiCUP removes duplicated reads, arising mainly from PCR library amplification, and creates a final bam file containing valid di-tags. After detecting a ligation junction, the conventional HiCUP pipeline truncates and discards any read spanning over it, considering one pair of interacting fragments per di-tag processed. However, in order to increase the valid di-tag reads (see **Chapter 3**), PECHi-C data were aligned using the newly devised **HiCUP combinations pipeline** (available on Github). Briefly, the new version splits all sequencing reads at each recognized ligation junction and retains all the resulting pair-wise combinations of interacting fragments within across the di-tags. All the resulting combinations are then put through an additional filtering step to retain only

valid and unique di-tags.

To detect "on target" di-tags, representative of interaction pairs involving specifically PEs, the final bam file produced by HiCUP was processed by the `get_captured_reads` script, which is provided with HiCUP. Briefly, the script compares the bam files against the .baitmap file, which includes the coordinates of the *DpnII* restriction fragments captured by the RNA baits, and identifies reads specifically mapping to the captured *DpnII* fragments, compiling a final captured.bam file which contains reads that exclusively map to the regions on interest. The `get_captured_reads` script was also used to calculated the final capture efficiency for PECHi-C data.

2.2.2.2 Capture Hi-C Analysis of Genome Organization (CHiCAGO) interactions calling

Combining two biological replicates for each timepoint of the naïve-to-primed transition, PECHi-C significant interactions were identified using the CHiCAGO pipeline [432]. CHiCAGO statistical model takes into account both technical and biological background components and uses bespoke normalization and multiple testing correction to identify significant interactions. For the processing of PECHi-C, adjacent *DpnII* restriction fragments were grouped into 5kb bins, excluding baited fragments, and CHiCAGO was run using the default parameters. For the multiple testing correction and p-value weighting, coefficients to adjust p-value weighting were identified with the `fitDistCurve.sh` script provided by the CHiCAGO tool suite, which determines the value of the coefficients based on the true positive rate for a given interaction distance in biological replicates. For PECHi-C the weights coefficients computed and used for multiple testing correction were: α = 52.3411565086191, β = -3.98123407981317 , γ = -17.2172007666019, δ = -7.01618020233379. CHiCAGO then assigns a score to each interaction, which represent the -log of weighted p-value, and a score $\geq$5 was used to identify PECHi-C significant interactions, according with previous reports that the threshold at score 5 maximises the enrichment of interacting regions for regulatory chromatin marks [433, 432].

2.2.3 Dimensionality reduction analyses

Principal Component Analysis (PCA) was performed on arcisine transformed CHiCAGO score to partition the single timepoints of the naïve-to-primed transition using the `prcomp` R method and `fviz_pca` function from *factoextra* R package [434], v1.0.7. Likewise, hierar-

chical clustering analysis was carried out using the `hclust` R function, using the complete linkage method which calculates the maximum distance between clusters.

K-means clustering approach was used to partition PECHi-C contacts using arcsine transformed CHiCAGO scores. The appropriate number of k was selected by applying the *elbow method*, calculating the within-cluster sum of squares (WCSS) in an iterative manner for k values between 2 and 10. The optimal number of k is represented by the values at which the WCSS abruptly decreases. A k = 7 was applied to imputed PECHi-C data and k = 8 was applied to non-imputed PECHi-C data. After identifying the more appropriate number of ks, k-means clustering was performed on arcsine transformed CHiCAGO scores using the `kmeans` R method and visualized using `pheatmap` from the *pheatmap* R package, v1.0.12 [435].

2.2.4 Imputation strategy of missing data points

For each timepoint, the CHiCAGO-detected list of interactions was implemented with a list of candidate PE-PE contacts and PE-Transcription Start Site (TSS) contacts with a distance range between 1kb and 10kb. The distance range was determined by plotting the contact's distance range distribution of CHiCAGO-detected interactions for each timepoint and choosing as cutoff the distance at which the number of interactions sharply decreased. A list of all annotated TSSs from protein coding genes was downloaded from Gencode database [436]. The list of candidate contacts was compiled using `bedtools window` function with the following parameters: `bedtools window -w 10000` and it was then implemented into the peak list of interactions identified by CHiCAGO [437]. After compiling the implemented list of contacts, a number of reads N was assigned to each contact pair. Specifically, for candidate contacts also identified by CHiCAGO, the imputed value of N was given by:

$$N = pmax(\frac{N}{(s_j * s_i)}, \frac{(Bmean + Tmean)}{(s_j * s_i)}) \tag{2.1}$$

where

- `pmax` = R method that returns the maximum of the given values
- N = number of observed read pairs spanning from other end (i) to baits (j)
- s_j/s_i = bait fragment-specific bias/other end fragment-specific bias
- Bmean/Tmean = brownian collision noise component/technical noise component

For candidate contacts which were not identified by CHiCAGO, the value of N was computed using the `distFun` from the Chicago R package [432] with the parameter's values computed by CHiCAGO for interactions with matching distance. Imputed N values were then normalized by a sample-specific scaling factor s_k computed by CHiCAGO that uses a similar strategy to that used in DESeq [432, 438].

2.2.5 Definition of PECHi-C interaction classes

Spearman's correlation was calculated for arcsine transformed CHiCAGO interaction scores using `cor` R function, `method` = "spearman". UP interaction class was defined applying a cutoff of $\rho \geq 0.4$; DOWN interaction class was defined with a cutoff of $\rho \leq$ -0.4, while interactions with -0.2 $\geq \rho \leq 0.2$ were included into the CONSTANT interaction class. At this stage, for the definition of the interaction classes, differentiated timepoints were excluded and only contacts with a CHiCAGO score ≥ 5 in at least one of the timepoints of the naïve-to-primed transition were included.

2.2.6 Gene Ontology enrichment analysis

Gene Ontology enrichment term analysis based on "biological processes" was performed using *clusterProfiler* R Bioconductor package, v4.2.0 [439]. To determine whether a gene ontology term was significantly over-represented, a p-value cutoff ≤ 0.05 was applied and Benjamini-Hochberg method was used for multiple testing correction.

2.2.7 Cut&Tag data processing

In collaboration with Babraham Bioinformatics (Babraham Institute, Cambridge), Cut&Tag sequencing data were aligned to GRCh38 reference genome using bowtie2 and a pileup of read counts was generated for the *DpnII* digested genome using `bedtools coverage` [427, 437]. Read counts were then normalized using DESeq2 [440] normalization scaling factor and normalized reads for three biological replicates were combined.

2.2.8 Linear regression model

C&T read counts were transformed using the variance stabilizing transformation function `vst` from *varistran* R package (counts for DE and NE samples were excluded from this analysis). An initial linear regression model was performed using `lm` R method and setting H3K27me3 and H3K4me1 read counts as response variable ($\hat{y}$) and taking a sum of

the predictor variables defined as time (t) and Spearman's classes (c_1 = UP; c_2 = DOWN). In the initial model, Spearman's interaction classes were modeled as dummy variables, whereas time was modelled as a continuous variable:

$$\hat{y} = \beta_0 + \beta_1 t + \beta_2 c_1 + \beta_3 c_2 + \beta_4 t c_1 + \beta_5 t c_2 + \epsilon \tag{2.2}$$

The initial model was further refined to include time as a discrete variable in order to account for the effect of Spearman's classes on H3K4me1 and H3K27me3 levels at specific timepoints rather than across the whole time course (k = coefficient number):

$$\hat{y} = \beta_0 + \sum_{i=1}^{8} \beta_k t_i + \sum_{n=1}^{2} \beta_k c_n + \sum_{i=1}^{8} \sum_{n=1}^{2} \beta_k t_i c_n + \epsilon \tag{2.3}$$

For visualization the function `allEffects` from *effects* R package was used, v.4.2.1 [441].

2.2.9 Motif Discovery analysis and transcription factors affinities

PE DNA sequences were computed using `getSeq` function of the *BSgenome* R package, v1.62.0 [442] and Position Weight Matrices (PWMs) for all annotated transcription factors (TFs) were downloaded from the HocoMoco database [443]. The HoCoMoco collection (version 11) contains 1302 mononucleotide PWMs for 680 human transcription factors, derived by systematic processing of more than five thousand ChIP-Seq experiments [443]. PWMs were then used to compute binding affinities of PE DNA sequences for any given TF using *tRap* R package, v0.7 [444]. One of the advantages of TRAP method over the other "hit" based methodology is that it does not rely on setting a specific threshold, but all the positions in the sequence contribute towards the calculation of the overall affinity, including low-affinity positions [444]. However, differences in DNA accessibility of the regions of interest were not accounted for in this analysis.

Kruskal-Wallis test was then applied to identify TFs with significantly different affinities between PEs of different interaction classes and only TFs with a significant difference in affinity for at least one of the three classes were selected, for a total of 113 hits (see **Appendix D**). A post-hoc Dunn's multiple comparison test was then computed to identify the specific classes showing significant differences for the 113 TF hits identified.

De novo motif discovery was then carried out using *rGADEM* R package, v2.42.0 [445]. Briefly, GADEM is an algorithm that couples the guided formation of space dyads with a

expectation-maximization (EM) algorithm [446]. After identifying over-represented sequences (lengths 3-6) that are used as "seeds", GADEM then uses a Genetic Algorithm (GA) to guide the formation of the space dyads from the identified seeds and returns PWMs for the inferred most represented motifs. The resulting PWMs were then compared, through Pearson's correlation, with PWMs of the 113 TFs that showed a significant affinity differences identified by the previous analysis. A lenient cutoff was applied and only TFs with a r $\geq$0.3 for at least one of *de novo* motifs identified were selected, identifying 63 candidate TFs.

HOMER was also used to carry out motif discovery analysis [447]. HOMER is a differential motif discovery algorithm that uses a zero or one occurrence per sequence scoring system coupled with hypergeometric or binomial enrichment calculations to determine motif enrichment. To perform differential motif discovery, HOMER uses control regions as background and it usually selects a total number of background regions of 50,000 or twice the total number of peaks provided. HOMER then performs a known motif enrichment analysis and a *de novo* motif discovery analysis. For the former, it screens its library of known motif against the background regions and the regions of interest and return motif with a p-value $\leq$ 0.05. For *de novo* motif discovery HOMER looks for motifs of length 8bp, 10bp and 12bp by default and it calculates the enrichment for specific sequences using a cumulative hypergeometric distribution or a cumulative binomial distribution. Motif discovery analysis was carried out for PE regions within the different interaction classes using the following parameters: `findMotifsGenome.pl hg38 -size given -mask`.

2.2.10 RNA-seq analysis

RNA-seq counts for hNES1 and H9 hESCs were downloaded from Gene Expression Omnibus (GEO accession: GSE123055) [424]. Specifically, RNA-seq counts were obtained for seven different timepoints during the naïve-to-primed transition: naïve (day 0), day 1, day 2, day 3, day 7, day 10, primed and differential expression analysis was carried out using the wrapper `DESeq` function from the *DESeq2* Bioconductor R package, v.1.34.0 [440]. In brief, DESeq2 normalizes raw counts using a computed scaling factor to account for differences in library size. After estimating the gene-wise dispersion to model the normalized read counts, it then fits a negative binomial model and performs Wald test of likelihood odd ratio for hypothesis testing. Normalized counts were extracted using the *DESeq2* `counts` function and z-scores were computed with the `scale` R function. The

`pheatmap` function from *pheatmap* R package, v.1.0.12 was used for visualization [435].

2.2.11 Chi-squared (χ^2) test of independence

CGIs coordinates were downloaded from UCSC Genome Browser [448]. The overlap between PE regions and CGIs, selecting CGIs regions within a 500 bp window of PEs, was computed with `findOverlaps` from the *GenomicRanges* R Bioconductor package, v1.46.0 [449]. To assess the association between PEs and CGIs, *regioneR* R Bioconductor package, v1.26.1. *regioneR* provides the tools to statistically evaluate the associations between region sets by performing permutation tests [450]. The function `overlapPermTest` was used to test if PEs within a given interaction class overlapped with CGIs more than expected, setting `ntime` = 1000. To then compare the association between PEs and CGIs across the three interaction classes a χ^2 test of independence was performed using the `chiquared` R method. To correct for overdispersion, which could result from incomplete independence of the variables, χ^2 tests was performed on randomly sampled CpG-positive (i.e. PEs overlapping with CGIs) and CpG-negative (i.e. PEs that do not overlap with any CGI) regions for a total of 100 times (same number of region were selected for both CpG-positive and CpG-neative regions). The standard deviation (sd) of the χ^2 values resulting from the 100 tests was then used to calculated the new value of degree of freedom (df) as follows:

$$df = \frac{sd^2}{2} \tag{2.4}$$

given the relationship between sd and df:

$$sd = \sqrt[2]{df} \tag{2.5}$$

The new value of df was then used to compute the adjusted p-value using the `pchisq` R method.

Following the same approach, χ^2 test of independence was used to probe the association between interaction classes and canonical PEs (coordinates of hESCs canonical PEs were obtained from Pachano, T., et al., 2021 [265]).

2.2.12 Log odds ratio and Empirical Cumulative Distribution Function

DPPA2/4 ChIP-seq bam files were provided by Rugg-Gunn lab (Babraham Insitute, Cambridge UK). Peaks were called using Macs2 peak-caller applying a cutoff of p-value = 10^{-2}

as follows: `macs2 callpeak -g hs -p 0.01 --nomodel --shift 0 --extsize 200/300 -B --SPMR --keep-dup all` [428]. Read count pileups for *DpnII* digested genome were generated using `bedtools coverage` [437]. The overlap between PE baited regions and DPPA2/4 peak regions was computed using `findOverlaps` from the *GenomicRanges* R Bioconductor package, v1.46.0, requesting a minimum overlap between regions of 10bp [449]. Fisher test was computed using the `fisher.test` R method and natural logarithm (log) of the odd ratio was visualized using `pheatmap` from the *pheatmap* R package, v1.0.12 [435].

DPPA2/4 ChIP-seq read counts were normalized using DESeq2 [440] normalization scaling factor to take into account the difference in library size and normalized reads for two biological replicates were combined. Empirical Cumulative Distribution Function was computed using `stat_ecdf` function from *ggplot2* R Bioconductor package, v3.3.5, with number of steps n = 5 [451].

3 Refinement of low cell number Capture-HiC for its use in human embryonic stem cells (hESCs)

3.1 INTRODUCTION

Hi-C allows genome-wide investigation of 3D chromatin structure and CHi-C enables the fine mapping of interactions of interest, which would normally require between 20 to 50 fold greater sequencing depth than traditional Hi-C libraries. The CHi-C protocol generally involves three main stages: Hi-C library generation, "capture" of regions of interest through hybridization of biotin-labelled RNA probes of the Hi-C library and identification of significant interactions via downstream computational analyses. However, this method typically requires large number of cells (30-40 million cells), which makes the protocol inaccessible when working with rare cell types, as for example in the specific case of cells from the early stages of organism development (e.g. naïve hESCs).

To overcome these limitations, our lab previously developed a CHi-C protocol that allows for a smaller number of cells as starting material [452]. In addition, the protocol makes use of a four-cutter enzyme, *DpnII*, as opposed to the more commonly used six-cutter enzyme *HindIII*, to increase the final resolution of CHi-C data. One of the major differences compared to current protocols is the use of Tn5-transposase for double-stranded DNA tagmentation and simultaneous insertion of sequencing adapters which largely reduces the timing of Hi-C and CHi-C library generation (**Figure 3.1** for a schematic of the protocol) [452].

This chapter presents the optimization of crucial steps of the previously developed low cell number CHi-C protocol [452] to robustly generate high-quality Hi-C and CHi-C library in hESCs. Additionally, it introduces the fine tuning of CHiCAGO (see **Methods**),

for the downstream analysis and detection of significant interactions in CHi-C data generated using a four-cutter restriction enzyme ([433]).

In collaboration with Ray-Jones H., Malysheva V. and Spivakov M., I first compared the performance of our *in house* developed CHi-C protocol and reagents with commercially available ones. I then optimized protocol parameters aiming to increase Hi-C and CHi-C final quality and introduced adjustments to optimally perform the protocol in hESCs.

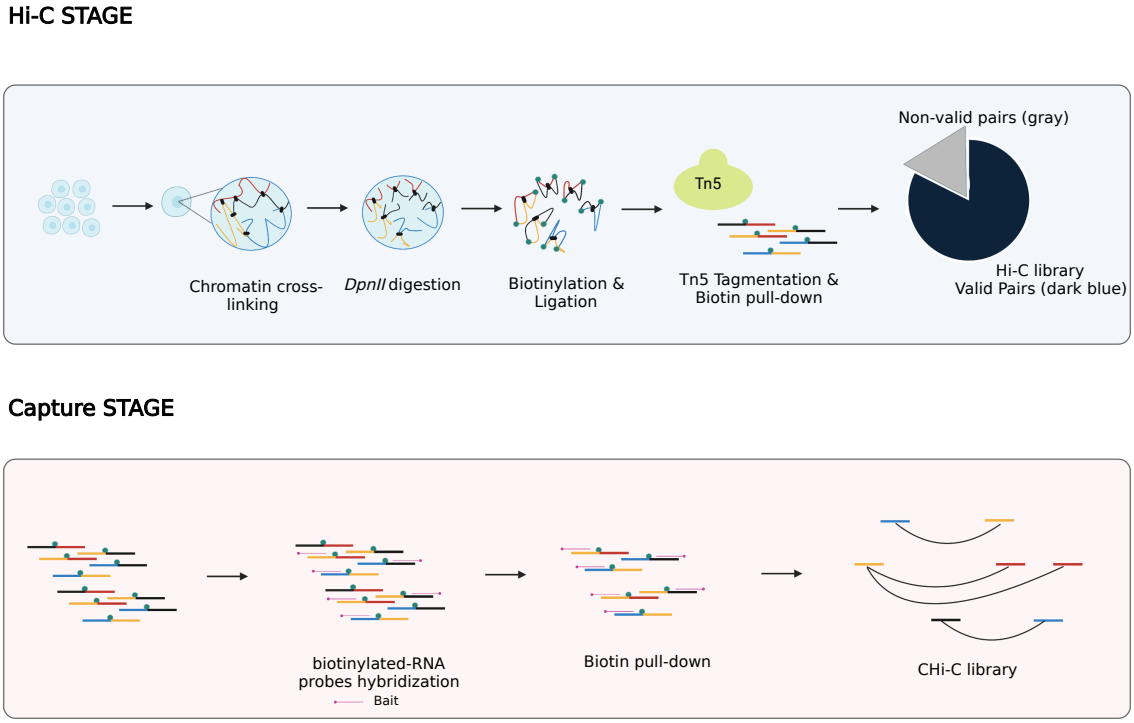


Figure 3.1: **Schematic of the Capture Hi-C methodology.** In the **Hi-C STAGE** of the protocol, following chromatin crosslinking and cell permeabilization, chromatin is digested using the four-cutter *DpnII* restriction enzyme. Restriction fragments are then marked with biotin prior to ligation and ligated Hi-C DNA is then processed for Tn5-mediated tagmentation with the parallel insertion of sequencing adaptors. Biotin-streptavidin pulldown allows for the enrichment of valid ligation products, represented by the dark blue portion of the pie chart, representative of "true" interacting chromatin regions. In the **Capture STAGE**, Hi-C libraries are hybridized with biotinylated RNA probes complementary to regions of interest. Biotin-streptavidin pulldown allows for the enrichment of interaction pairs which involve, at least at one end, fragments of interest.

3.2 RESULTS

3.2.1 Optimizing Tn5-mediated tagmentation reaction for the generation of Hi-C libraries

The first stage of the protocol, the Hi-C stage, involves the use of Tagmentation Ready Tn5-transposase (henceforth referred to as TR-Tn5) for double-stranded DNA tagmentation and parallel insertion of sequencing adapters (**Figure 3.1**). The substantially large quantity of Tn5 required to perform the protocol for large-scale sequencing projects represented a severe limitation for its application, due to the cost of commercially available Tn5. Therefore, following a published procedure [453] previously adopted by Malysheva V., I produced TR-Tn5 transposase and I compared TR-Tn5 with commercially available Tn5 enzyme (henceforth referred to as TDE) in order to generate high-quality Hi-C libraries.

I first compared different tagmentation reaction conditions in monocyte cells (used for the initial establishment of the miniaturized CHi-C protocol) using TDE and TR-Tn5. I assessed the final Hi-C library quality generated using TDE or TR-Tn5 by testing the following parameters: different glycerol levels in the final volume of the tagmentation reaction (glycerol levels are here defined as $< 5\%$ or $> 5\%$) and different ratios of AmpureXP beads to DNA (AmpureXP:DNA) for final DNA extraction after PCR amplification (conditions summarized in **Table 3.1**).

Protocol Step				
Glycerol levels	$< 5\%$	$< 5\%$	$> 5\%$	$> 5\%$
AmpureXP:DNA ratio	1X	0.7X	1X	0.7X

Table 3.1: **Summary of glycerol level percentages and AmpureXP:DNA ratios tested.** The table summarizes the glycerol levels of the final Tn5-mediated tagmentation reaction tested for both TDE and TR-Tn5 and the AmpureXP:DNA ratio for the final Hi-C DNA extraction after PCR library amplification.

To assess the final quality of HiC libraries I measured the final percentage of valid pairs present, usually aiming for a high percentage of valid pairs (typically between 65%-95%, **Figure 3.1** Hi-C library pie chart in top panel) to minimize the proportion of uninformative ligation products, hence to maximize the number of useful sequencing reads for the

downstream analysis.

Our results showed that, while TDE performed marginally better than TR-Tn5 (TR-Tn5 yielded 60-50% valid-pairs while TDE showed percentages >70% reaching, in some cases 90%), they were both able to generate Hi-C libraries with a valid-pair percentage ranging from 60% to 90% (see **Figure 3.2, A**).

Varying glycerol levels of the tagmentation reaction and AmpureXP:DNA ratio affected the final quality of the Hi-C library. Although glycerol represents a stabilizing agent for the storage of enzymes, it can interfere with enzymatic activity if present at high concentrations in the final reaction volume. More specifically, in our case levels of glycerol $< 5\%$ of the final reaction volume always resulted in >60% of valid-pairs, dropping to 50% or below for glycerol levels $> 5\%$ (**Figure 3.2, B**). I concluded that levels above 5% lead to sub-optimal tagmentation levels, which becomes evident only after the sequencing stage of the HiC library. Therefore, it is crucial to control for glycerol levels in order to generate a final high percentage of valid pairs. In addition, a 0.7X AmpureXP:DNA ratio also generally yielded better quality Hi-C libraries, showing a percentage of valid pairs >70% in nearly all cases (**Figure 3.2, C**). A 0.7X AmpureXP:DNA ratio allows to remove short fragments from the HiC library that could potentially results from the random insertion of adapters by Tn5-transposase or through the previous *DpnII* digestion step. These might not be representative of "true" interactions, leading to a decrease in percentage of valid pairs if retained in the final HiC library. In addition, the presence of this specific population of short fragments would only become clear after the sequencing and the data processing stage, therefore establishing the correct AmpureXP:DNA ratio was essential to robustly generate good quality final libraries.

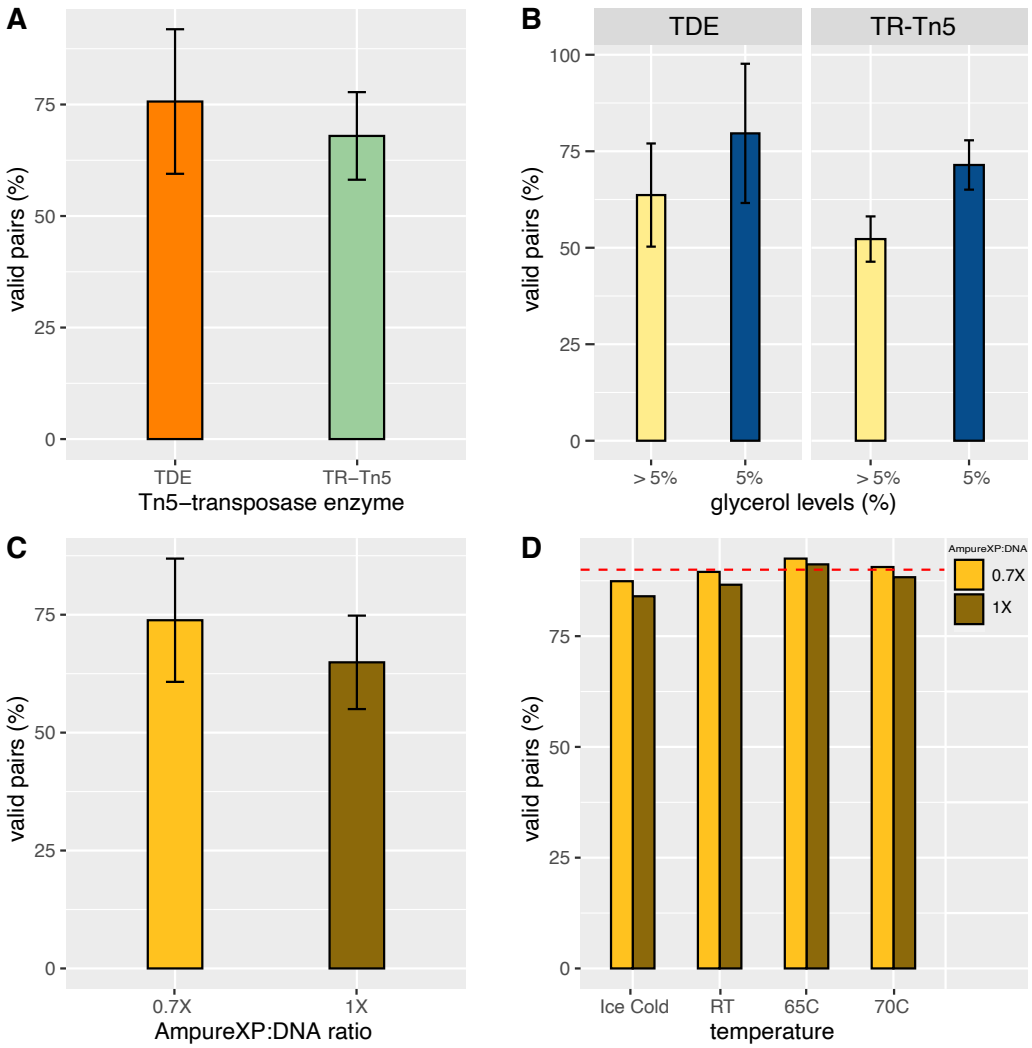


Figure 3.2: **Optimization of Tn-5 tagmentation reaction conditions for generation of good quality Hi-C and CHi-C libraries.** Hi-C libraries were generated for monocyte cells, primed (H9) and naïve (hNES1) hESCs starting from 1x10^6 cells. **A.** Barplots showing the percentage of valid pairs in the final Hi-C library (y-axis) following tagmentation with TDE enzyme (orange) and TR-Tn5 enzyme (green). **B.** Barplots showing impact of the Tn5-mediated tagmentation reaction's glycerol levels (>5 yellow; ≤ 5 blue) on final percentage of valid pairs (y-axis) in the final Hi-C library for both TDE enzyme and TR-Tn5 enzyme. **C.** Barplots showing the impact of the ratio of AmpureXP beads to DNA (1X, gold and 0.7X, brown) of the final DNA extraction on the final quality of Hi-C libries (represented as percentage of valid pairs, y-axis). A ratio of 0.7X resulted in higher percentage of valid pairs (gold), compared to 1X ratio (brown), in some cases reaching $\geq 80\%$ of valid pairs. **D.** Barplots showing the effect of the increased temperature of post biotin-streptavidin pulldown (x-axis) on the percentage of valid pairs in the final Hi-C library (y-axis). Washes at 65ºC washes yielded the best final percentage of valid pairs and the combination of post biotin-streptavidin pulldown hot washes and 0.7X AmpureXP:DNA ratio resulted in $\geq 90\%$ valid pairs in the final Hi-C library (red dashed line). For all Barplots, error bars represent standard deviation.

To further improve the final percentage of valid pairs, I sought to reduce spurious biotin-streptavidin association events and increase the ratio of "true" interaction pairs in the final library. Our rationale was that by increasing the temperature I could increase the stringency of post biotin-streptavidin pulldown washes, hence reducing the presence of non-informative ligated pairs in the final Hi-C library (i.e. higher percentage of valid pairs). I set out to test four different temperature conditions: standard RT, 65ºC, 70ºC and ice-cold washes (henceforth "hot washes" will generally refer to both 65ºC and 70ºC washes and "cold washes" will refer to ice-cold conditions).

Our results showed that while "cold washes" always resulted in a worse percentage of valid pairs in the final HiC library, as expected, "hot washes" improved the overall quality of the resulting library with an increase of at least 5% in the final valid-pairs percentage. Particularly, washes at 65ºC show the best improvement, with percentage of valid-pairs >90% (**Figure 3.2, D**).

In summary, I concluded that glycerol percentage of the final tagmentation reaction volume should never exceed 5%. A ratio of AmpureXP beads:DNA of 0.7X improves the percentage of valid-pairs through the removal of very short DNA fragments which are likely the result of the insertion of sequencing adapters within the same restriction fragment, therefore not representative of true interacting pairs. Likewise, higher stringency washes post biotin-streptavidin pulldown performed at 65ºC show an improvement of at least 5% of valid-pairs in the final library, resulting in many cases in >90%.

3.2.2 Refining protocol conditions for the generation of Capture Hi-C libraries in hESCs

Although the previously described conditions consistently generated Hi-C libraries with a high final percentage of valid pairs, I noticed that Hi-C and CHi-C libraries generated in primed and naïve hESCs consistently under performed when compared to libraries generated in monocytes (**Figure 3.3**).

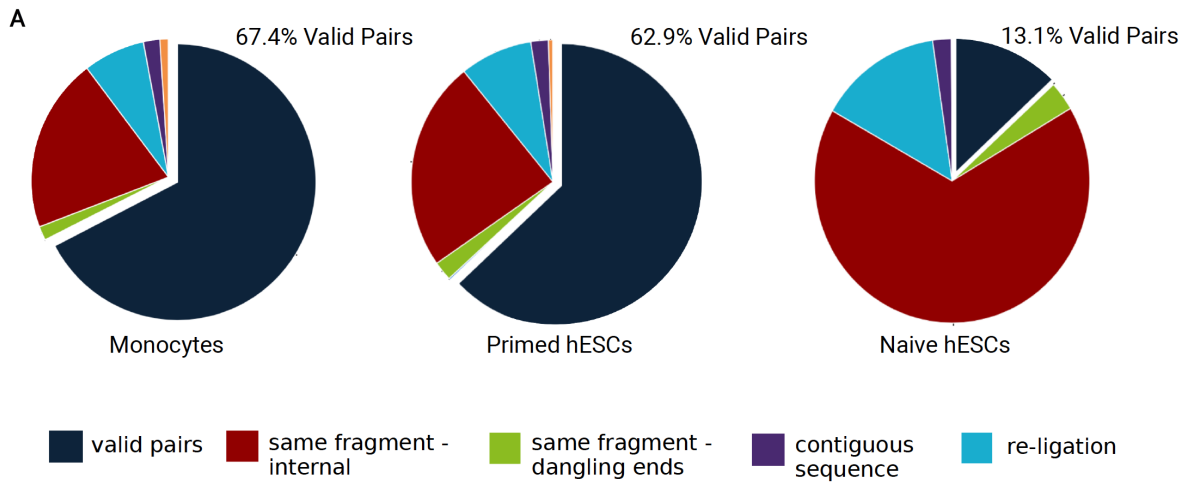


Figure 3.3: **Protocol for generating Hi-C and CHi-C libraries significantly under performed for primed and naïve hESCs.** Capture Hi-C libraries were generated for monocyte cells, primed (H9) and naïve (hNES1) hESCs starting from 1x10^6 cells. Pie chart generated by the HICUP pipeline [431] showing the percentage of valid pairs in the final CHi-C libraries generated for monocytes, primed hESCs and naïve hESCs (left to right). The optimized conditions used to generate good quality libraries for differentiated cells (monocytes, ~70%valid pairs) were not optimal for hESCs. Primed (H9) hESCs slightly under performed (middle pie chart, ~60% valid pairs) and naïve (hNES1) hESCs generated very poor quality Hi-C and CHi-C libraries, with very low percentage of valid pairs (~10% right hand pie chart) and a higher percentage of same-internal fragments (red portion of right hand pie chart, ≥50%).

Therefore, I set out to optimize conditions to generate good quality Hi-C and CHi-C libraries in primed (H9) and naïve (hNES1) hESCs in order to be able to perform CHi-C to investigate the emergence of PEs regulatory network in hESCs.

I compared conditions for both the initial Hi-C stage of our protocol and the second Capture stage across monocytes, primed (H9) and naïve (hNES1) hESCs (a schematic of the experimental setting is shown in **Figure 3.4**).

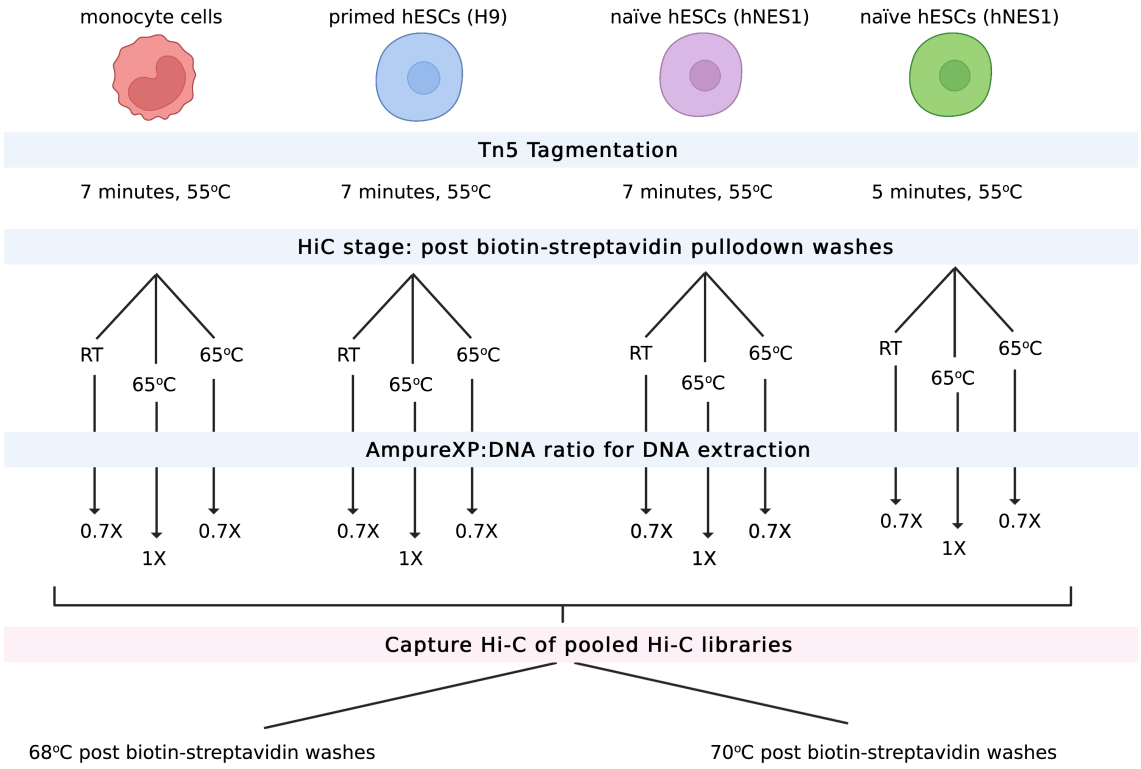


Figure 3.4: **Experimental design for the optimization for Hi-C and CHi-C library preparation in primed (H9) and naïve (hNES1) hESCs.** Hi-C DNA was generated for monocyte cells in addition to primed (H9) and naïve (hNES1) hESCs starting from 1x10^6 cells, and tagmented following the standard conditions: 7 minutes at 55ºC. For naïve cells an additional tagmentation reactions was carried out for 5 minutes at 55ºC alongside the standard conditions. For each sample, the tagmented Hi-C material was then split into three equal reactions to test three different Hi-C post biotin-streptavidin pulldowns coupled with different DNA extraction conditions after PCR amplification: room temperature washes (RT) + 0.7X AmpureXP:DNA ratio; 65ºC washes + 1X AmpureXP:DNA ratio and 65ºC washes + 0.7X AmpureXP:DNA ratio. Following DNA purification, Hi-C libraries were pooled for all samples. The resulting pool of Hi-C libraries was then split in two equal reactions to test two different conditions for the Capture Hi-C post biotin-streptavidin washes: 68ºC and 70ºC.

Firstly, our results confirmed what previously observed: "hot washes" at 65ºC resulted in a higher percentage of valid pairs compared to washes carried out at RT, yielding >65% in most cases (**Figure 3.5, A**). Notably, while primed (H9) hESCs and monocytes showed very comparable results, naïve (hNES1) hESCs under-performed for all conditions tested. Nevertheless, overall washes at 65ºC improved greatly the quality of the final HiC libraries even in the case of naïve (hNES1) hESCs with Hi-C libraries showing >50% valid-pairs, a ≥1.5 fold increase compared to RT washes (**Figure 3.5, A**). Likewise,

0.7X AmpureXP:DNA ratio for DNA extraction showed an increase in percentage of valid pairs for all samples, resulting in an increase of 5% or more as previously observed (interestingly, this was not the case for naïve cells samples that had gone through a tagmentation of 7 minutes as opposed to 5 minutes as shown in **Figure 3.5, B**). While driving an increase of the final Hi-C library quality for naïve and primed samples alike, AmpureXP:DNA ratio of 0.7X also resulted in a loss of material between 1.2 -1.5 fold in hESCs, considerably affecting the amount of DNA required for the subsequent Capture stage of the protocol. Therefore, we concluded that a 1X AmpureXP:DNA ratio should be applied when generating Hi-C libraries for hESCs, in order to limit the impact on the final yield of Hi-C library.

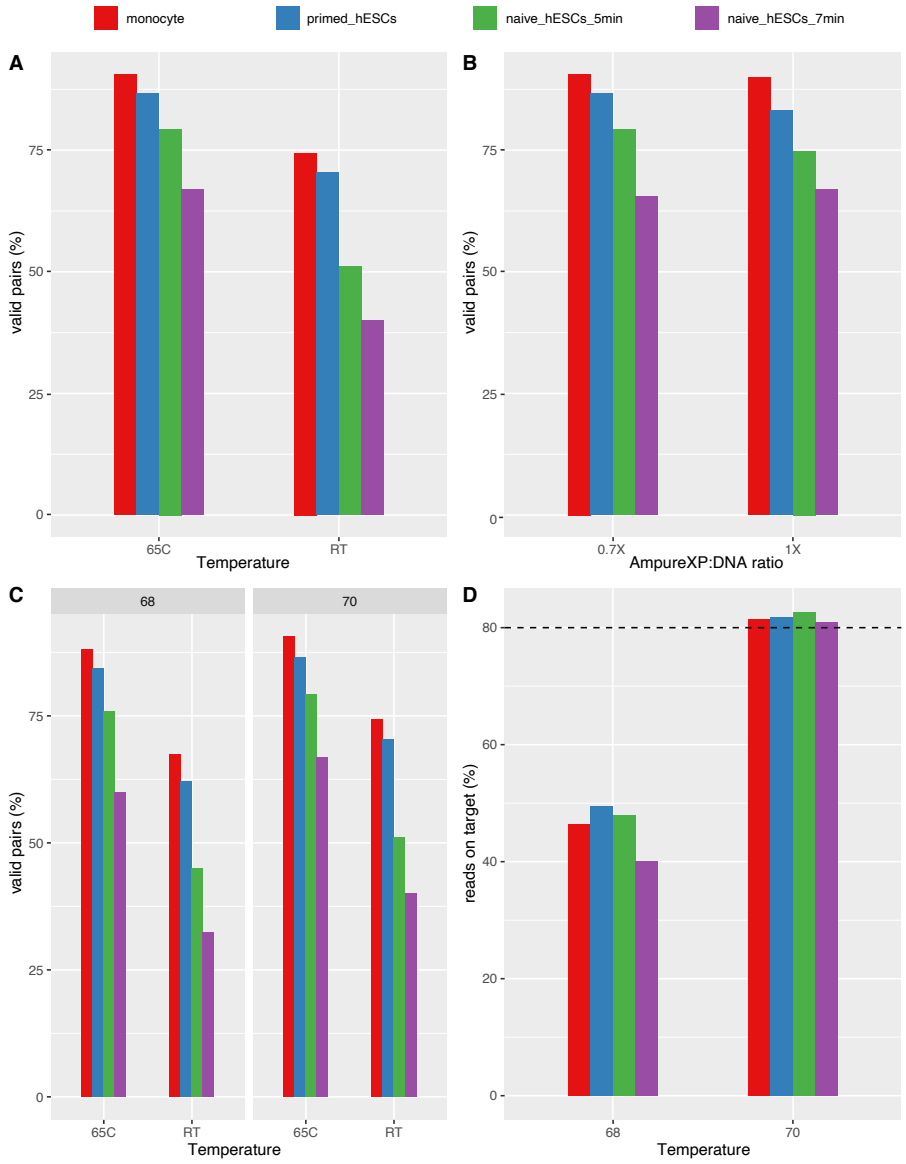


Figure 3.5: **Optimizing conditions for generation of good quality Hi-C and CHi-C libraries in primed (H9) and naïve (hNES1) hESCs.** Capture Hi-C libraries were generated for monocyte cells (red), primed H9 hESCs (blue) and naïve hNES1 hESCs (green: naïve hNES1 hESCs that have been tagmented for 5 minutes; purple: naïve hNES1 hESCs that have been tagmented for 7 minutes) starting from $1x10^6$ cells. **A.** Barplots showing the effect of 65ºC post biotin-streptavidin pulldown Hi-C washes compared to room temperature (RT) (x-axis) on the final percentage of valid pairs of Hi-C libraries (y-axis). **B.** Barplots showing the effect on the percentage of valid pairs (y-axis) of 1X AmpureXP:DNA ration compared to 0.7X AmpureXP:DNA ratio for final Hi-C DNA purification (x-axis). **C.** Barplots showing the effect of two different temperature conditions tested for the Capture Hi-C post biotin-streptavidn pulldown washes: 68ºC and 70ºC (left hand panel and right hand panel, respectively) on the final percentage of valid pairs (y-axis) in combination with two different temperatures for the Hi-C post biotin-streptavidn pulldown washes (x-axis). **D.** Barplots showing the effect on the percentage of reads on target (y-axis) of Capture Hi-C post biotin-streptavidin pulldown washes at 68ºC compared to 70ºC (x-axis).

Once the best conditions for the generation of Hi-C libraries in hESCs were identified, I set out to optimize conditions of the Capture stage of the protocol (**Figure 3.1**) in order to increase the proportion of contacts involving our regions of interest (i.e. PEs), hence the proportion of sequencing reads available, in the final CHi-C library. Similarly to what previously done for the Hi-C stage, I increased the stringency through increasing the temperature of post biotin-streptavidin pulldown washes that follow the hybridization of the baits, aiming to increase the final capture efficiency (defined as percentage of reads mapping to target regions in the final CHi-C library). I set out to test two different temperature conditions: the standard 68ºC and 70ºC. Ideally, increasing the temperature should result in the enrichment of the final CHi-C library for "on-target" products (i.e. fully hybridized baits to Hi-C ligation products) and minimize the presence "off-target" fragments (i.e. partially hybridized baits to Hi-C ligation products). As shown in **Figure 3.5, C**, washes at 70ºC significantly improved the percentage of reads "on-target" compared to the standard 68ºC in all samples, showing an increase of $\geq$1.5 fold in all cases. Moreover, I observed that the combination of 65ºC washes at the Hi-C stage with 70ºC washes at the Capture stage resulted, in most cases, in a better overall quality of the final CHi-C library, with both higher percentage of valid pairs, at 60% or above, and higher percentage of reads on target, between 60%-80%. (**Figure 3.5, D**).

Although 70ºC washes resulted in higher enrichment of our final CHi-C library, in some cases reaching $>$85%, the high temperature highly affected the final overall yield of the library, resulting in a loss of material of $\geq$2.3 fold. Low CHi-C library yields greatly affected the ability to reach the sequencing coverage necessary, hence resulting in severely under-powered CHi-C data.

Therefore, based on these results I determined that 68ºC represented the best condition for post biotin pulldown washes at the CHi-C stage of the protocol. This allows to achieve a capture efficiency that ranges between 20%-45% (i.e. 17-35 fold enrichment for interactions containing regions of interest) and it grants the final yield necessary to obtain appropriate sequencing coverage required for downstream analysis. In addition, to validate the robustness of detected interactions by CHiCAGO I computed the overlap of CHiCAGO detected interactions after performing a 20% down-sampling of the CHi-C data, as shown by the venn diagrams in **Supplementary Figure A2, Appendix A**. In all cases, the computed overlap after down-sampling ranged between 70% and 50%.

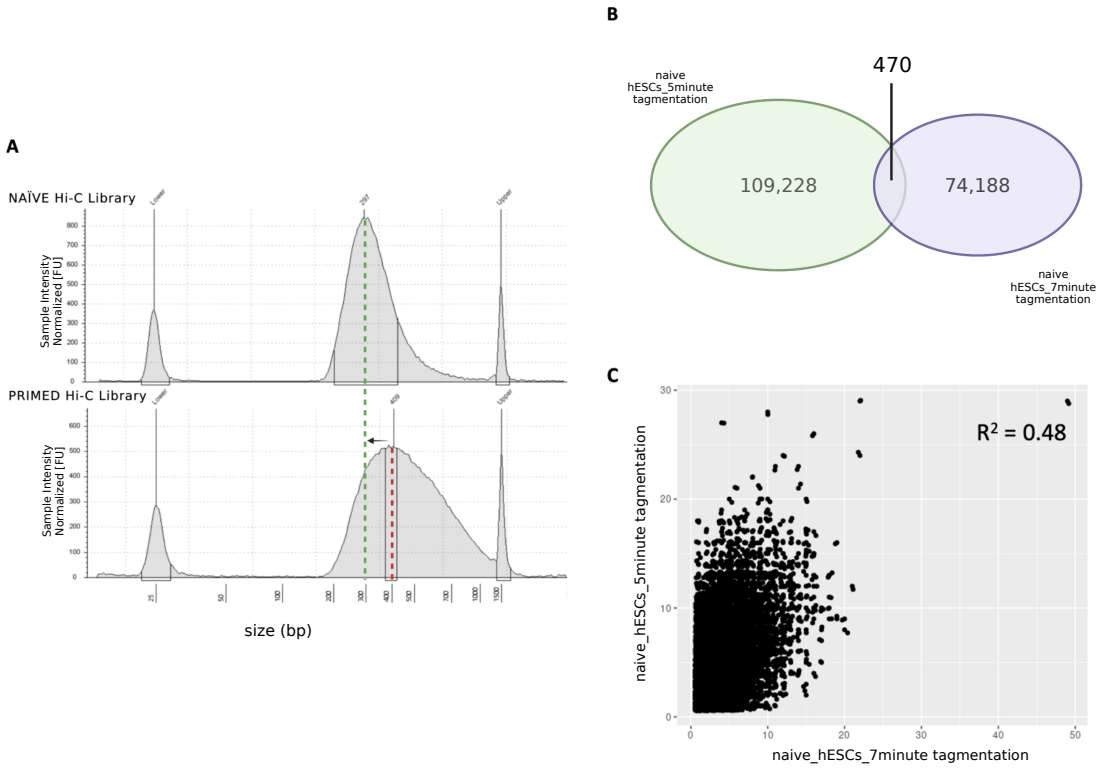


Figure 3.6: **Comparing Tn5-tagmentation conditions in naïve (hNES1) hESCs. A.** D1000-4200 Agilent Tapestation profiles of Hi-C libraries in naïve hNES1 (top) and primed H9 (bottom) hESCs. Naïve hNES1 show higher tagmentation levels compared to primed H9, with an average fragment size 2 times smaller than primed hESCs (∼200bp, marked by the green dashed line, and ∼400bp, marked by the red dashed line, respectively). **B.** Venn diagram of the overlap between contacts detected by CHiCAGO in naïve hNES1 after 5-minute tagmentation (green) compared to contacts detected in naïve hNES1 after 7-minute tagmentation (purple). **C.** Scatterplot showing the correlation between number of reads per unique bait captured in naïve hNES1 tagmented for 5 minutes (y-axis) and 7 minutes (x-axis), $R^2 = 0.48$.

As mentioned earlier, I observed that naïve cells, in particular, always under performed when compared to monocytes or primed cells, showing a lower percentage of valid pairs in the final Hi-C library (**Figure 3.3**). In addition, they consistently showed higher tagmentation levels when tagmented under standard conditions (i.e. 55ºC, 7 minutes), as shown in **Figure 3.6, A**), potentially due to higher DNA accessibility in these cells affecting the initial *DpnII* digestion and ultimately resulting in the presence of uninformative products in the final library. Therefore, I set out to test if reducing tagmentation time from the standard 7 minutes to 5 minutes could improve further the quality of Hi-C libraries for naïve hESCs by minimizing the potential over-tagmentation levels, hence the proportion of uninformative short fragments (**Figure 3.4**). Indeed, naïve sam-

ples tagmented for 5 minutes performed overall better compared to samples tagmented for 7 minutes, showing between 10% and 20% increase of percentage of valid pairs and higher percentage of reads on-target, reaching in some case a capture efficiency >80% (**Figure 3.5**). However, the two conditions showed a small overlap between the contacts detected and at same sequence coverage, the number of reads per unique bait showed poor correlation (R^2=0.48. **Figure 3.6 B and C**), which could be a consequence of the Tn5-mediated tagmentation. Indeed, it is plausible that the duration of the tagmentation reaction and the DNA:enzyme ratio could lead to a scenario whereby the Tn5 enzyme gains access to different cutting sites, hence generating qualitatively different fragments, ultimately resulting in different types of interactions being represented in the final Hi-C and CHi-C library. Based on these results, I decided against using different tagmentation conditions for primed and naïve cells and used tagmentation time set at 7 minutes for all samples processed. Nevertheless, as shown in **Figure 3.5, A**, these conditions in combination with 65ºC washes post biotin pulldown at the Hi-C stage still allowed us to reach a Hi-C library with a percentage of valid pairs ≥50% and a percentage of reads on target up to 40% (**Figure 3.5, C and D**).

3.2.3 Using *in house* buffers for the generation of Capture Hi-C libraries

In addition to the use of RT-Tn5 transposase produced in the lab, our protocol also makes use of *in house* buffers and conditions for the hybridization step at the Capture stage, which have been designed to minimize the cost of the protocol and to contain specific reagents tailored to the use of Tn5-mediated adapter insertion. I then set out to compare our *in house* protocol with the commercially available one (hereafter referred to Protocol A and Protocol B, respectively). Again, I compared between 68ºC and 70ºC CHi-C post biotin-pulldown washes and assessed the yield and the percentage of reads on target of the final CHi-C library (a schematic of the experimental design is shown in **Figure 3.7**).

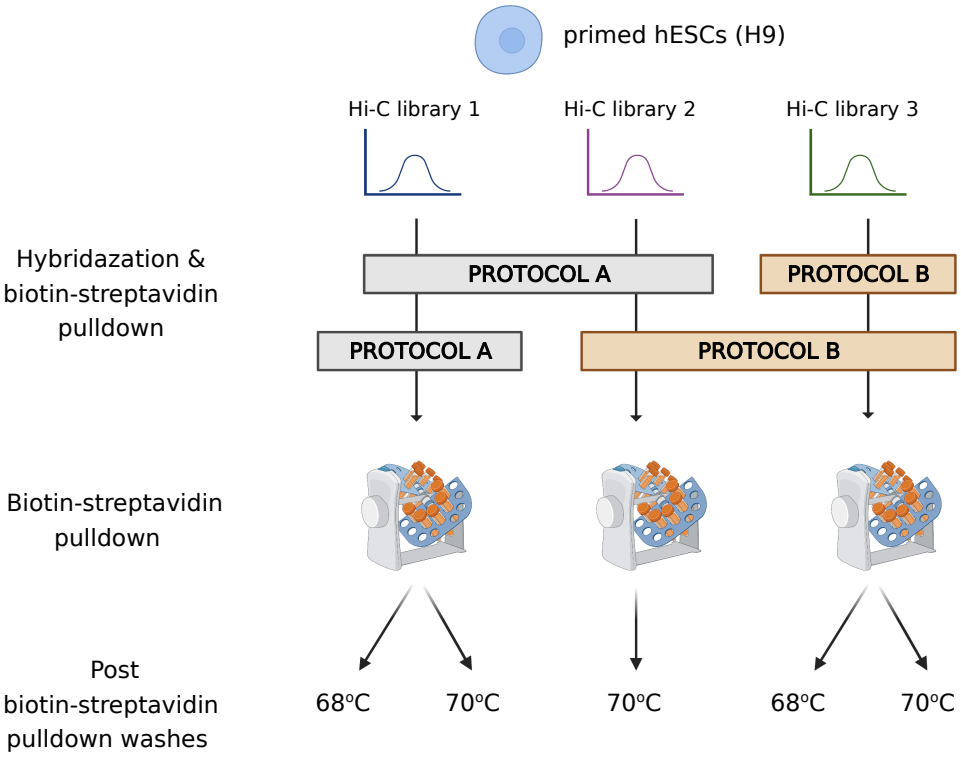


Figure 3.7: **Experimental plan comparing *in house* and commercially available protocols: schematic of the experimental plan.** Hi-C libraries were generated for primed H9 hESCs, starting from 1x10^6 cells. Each library was then hybridized with RNA-biotinylated probes following conditions specified by the *in house* protocol, (Protocol A, grey) or following the commercially available protocol's instructions (Protocol B, light brown). In addition, one of the samples was processed using a combination of commercially available buffers with *in house* protocol conditions (here represented by Hi-C library 2). Post biotin-streptavidin pulldown washes were performed at 68°C and 70°C to compare the final yield between protocols, except in the case of Hi-C library 2, for which post biotin-streptavidin pulldown washes were carried out at 70°C.

Our results confirmed that 70°C washes, although resulting in a higher capture efficiency for both Protocol A and Protocol B, ≥80% (**Figure 3.7**), considerably affected the final CHi-C library yield, in some cases resulting in a >5 fold reduction. While I did not observed any significant difference between Protocol A and Protocol B when I carried out washes at 70°C, always reaching a capture efficiency between 80%-90%, strikingly when washes were carried out at the standard temperature of 68°C, Protocol A performed better than Protocol B, generating libraries with a capture efficiency of 40% as opposed to the observed 10%, respectively (**Figure 3.8**).

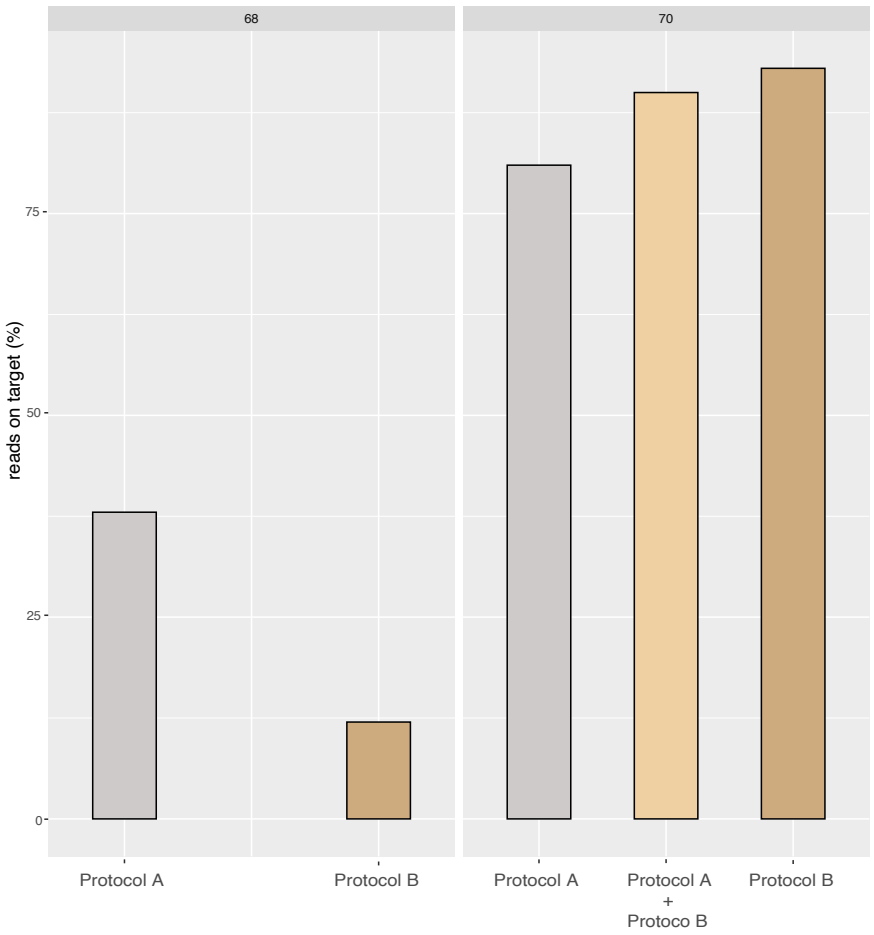


Figure 3.8: **Comparison of *in house* and commercially available protocols.** Hi-C libraries were generated for primed H9 hESCs, starting from 1x10^6 cells. Barplots comparing percentage of reads on target (y-axis) for CHi-C libraries processed following Protocol A (grey) and Protocol B (brown) performing post biotin-streptavidin Capture Hi-C washes at 68ºC (left panel) and 70ºC (right panel).

Here I identified optimal conditions for Capture stage using our *in house* protocol (i.e. Protocol A). Despite the high capture efficiencies obtained after the 70ºC washes, reaching in some cases $\geq 90\%$, high temperatures of post biotin pulldown washes at the Capture stage significantly affected the final yield of the resulting library and, consequently, the final sequencing coverage.

In summary, I established that for CHi-C experiments in hESCs the following condi-

tions should be applied when generating libraries through our *in house* protocol (**Table 3.2, Supplementary Figure A.1, Appendix A**):

Optimized Parameters	Tn5 Tagmentation	Hi-C stage	Capture stage
Glycerol content	<5%	/	/
Tagmentation temp.	55ºC	/	/
Tagmentation time	7 minutes	/	/
Washes Temp. post pulldown	/	65ºC	68ºC
AmpureXP:DNA ratio	/	1X	1X
Reagents	TR-Tn5/TDE	Protocol A	Protocol A

Table 3.2: Summary of the optimal conditions defined for the generation of Capture-HiC in hESCs

3.2.4 Increasing the yield of informative reads in four-cutter restriction enzyme derived (Capture) Hi-C

In order to increase Hi-C and CHi-C resolution, our protocol uses a four-cutter enzyme, *DpnII*, as opposed to the more commonly used six-cutter *HindIII*. One of the challenges that *DpnII* poses is a significant loss of HiC/CHi-C valid sequencing reads during HiCUP alignment and pairing (see **Methods**). The conventional HiCUP pipeline used for the processing of Hi-C and CHi-C libraries has been originally designed for the analysis of libraries produced using *HindIII* which generates 5kb long restriction fragments, while *DpnII* produces an average fragment size of 430bp. In addition, the use of tagmentation in our protocol, instead of sonication with subsequent fragment size-selection, results in fragments above 700bp being retained in the final libraries. This ultimately means that *DpnII*-derived Hi-C/CHi-C di-tags have properties that HiCUP is not tuned to process correctly. For instance, one of the major effects observed is reads spanning over the ligation junctions, which HiCUP normally truncates and filters out. In the case of ligation products containing more than two restriction fragments, the truncation and subsequent rejection of reads leads to a potential loss of valid read pairs, hence of informative di-tags representative of "true" interactions (**Figure 3.9**). Briefly, the conventional HiCUP pipeline truncates and discards any read that spans over a ligation junction, considering one pair of interacting fragments per di-tag processed (as shown in **Figure 3.9**, left branch of the diagram). The optimized version of the pipeline, **HiCUP combinations** re-

tains all the resulting pair-wise combinations of interacting fragments within and across the di-tags, after splitting sequencing reads at each recognized ligation junction, instead of discarding the truncated reads. All the resulting combinations are then put through an additional filtering step to retain only valid di-tags (as shown in **Figure 3.9**, right branch of the diagram)

Considering the substantial differences between the two protocols, in collaboration with HiCUP developer Steven Wingett, I devised a modified method, **HiCUP combinations**, tailored for the processing of *DpnII*-derived Hi-C/CHi-C libraries (**Figure 3.9**, see Methods **section 2.2.2.1**).

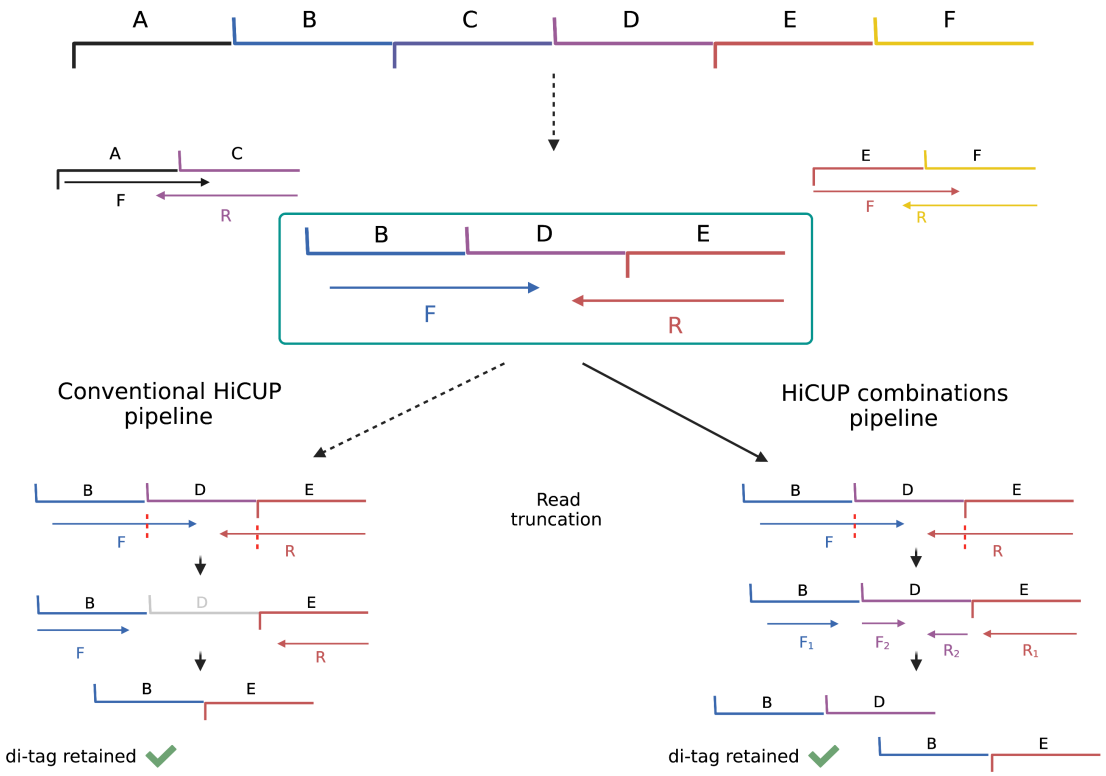
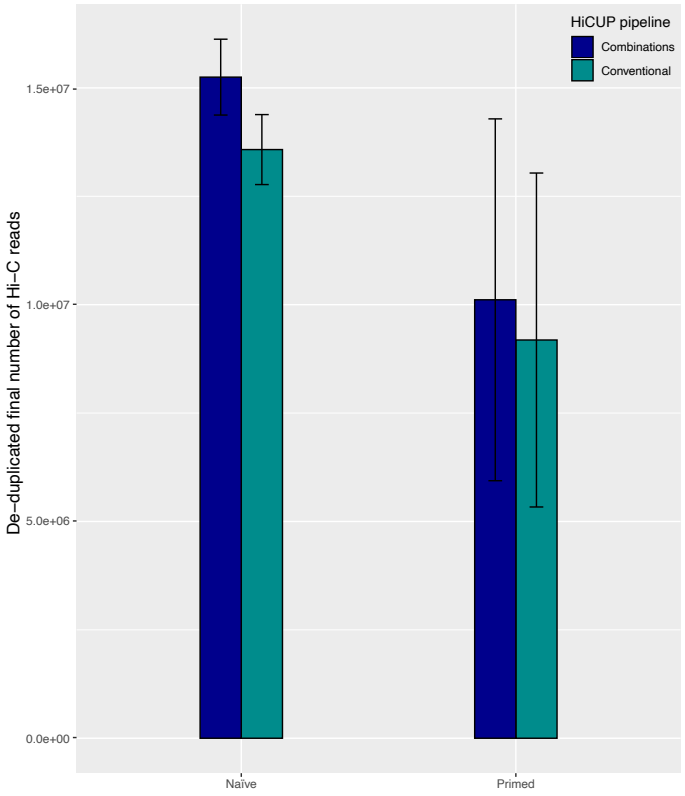


Figure 3.9: **Schematic of the processing of Hi-C sequencing reads by the HiCUP combinations pipeline.** When processing Hi-C paired-end (PE) sequencing reads, the conventional HiCUP pipeline truncates paired-end sequencing reads from the 5′ end at a ligation junction, generating individual single-end reads. The individual reads are then re-paired or discarded. HiCUP applies filters to identify and discard invalid di-tags that could result from scenarios such as same-fragment ligation, adjacent ligation and/or re-ligation events to retain only valid di-tags containing two fragments representative of ”true” interactions. The conventional HiCUP pipeline truncates and discards any read that spans over a ligation junction, considering one pair of interacting fragments per di-tag processed (as shown by the schematic of the left branch on the diagram). In the optimized version of the pipeline, **HiCUP combinations** retains all the resulting pair-wise combinations of interacting fragments within and across the di-tags instead of discarding the truncated reads, after splitting sequencing reads at each recognized ligation junction. All resulting combinations are then put through an additional filtering step to retain only valid di-tags (as shown by the right branch of the diagram).

To determine the proportion of reads rescued with the newly established HiCUP combinations pipeline, I compared the final number of reads of CHi-C libraries, generated in naïve and primed hESCs, aligned with the two different pipelines. For both samples I observed a 10% increase in unique final reads (**Figure 3.10**).



	Gained Hi-C reads (%)
Naive hESCs	10.96%
Primed hESCs	9.17%

Figure 3.10: **The HiCUP combinations pipeline retrieves Hi-C sequencing reads for naïve hNES1 and primed H9 hESCs.** Barplots comparing Hi-C libraries for naïve hNES1 and primed H9 hESCs for two biological replicates that have been aligned and processed using the HiCUP conventional pipeline, cyan (retrieving 13,577,498.5 and 9,185,315 de-duplicated reads, respectively) or the HiCUP combinations pipeline, blue (retrieving 15,248,534 and 10,112,634, respectively). The alignment and the pre-processing of Hi-C sequencing reads using the HiCUP combinations pipeline resulted in $\sim$10% gain of final de-duplicated Hi-C sequencing reads for both samples. Error bars represent standard deviation.

Indeed, these results confirmed that the HiCUP combinations pipeline allows to rescue up to 10% of valid sequencing pairs from *DpnII*-derived libraries in both naïve and primed hESCs.

3.2.5 CHiCAGO optimization for four-cutter restriction enzyme

In collaboration with Malysheva V., Ray-Jones H. and Spivakov M., I sought to compare CHi-C data produced with *HindIII* or *DpnII* and standardize CHiCAGO parameters for the analysis of CHi-C data generated with a four-cutter restriction enzyme [433].

3.2.5.1 CHiCAGO background estimation for four-cutter restriction enzyme

As shown in **Figure 3.11, A**, four-cutter enzymes preferentially detected shorter range interactions (most significant interactions detected within a 3kb-100 kb distance range) compared to six-cutter data (most significant interactions detected within a 100kb-400kb distance range). One of the reasons for this difference might lie in the inherent higher sparsity of data generated with a four-cutter enzyme. I devised a binning strategy aiming to mitigate this effect (**Figure 3.11, B**, bottom left panel) and I compared the distance range of interactions detected for 5kb-binned data and *HindIII*-derived data. Briefly, adjacent *DpnII*-derived restriction fragments are grouped in 5kb bins, with the possibility to include or exclude the "baited" restriction fragments. Excluding the "baited" fragments from the 5kb bins allows to retain the higher resolution given by *DpnII* at the captured regions of interest.

As shown in **Figure 3.11, B**, the binning resulted in the increase of the distance range of interactions to 30kb-200kb. But while binning improves the overlap between interactions detected in four-cutter data compared to six-cutter data, it didn't show a complete rescue. I then checked if this difference in distance range also reflected in the ability to detect biological relevant interactions. Our results showed that, in both cases, the other-ends of the interactions detected were enriched for biologically relevant histone marks (i.e. H3K27me3, H3K4me3, H3K36me3, H3K27ac) suggesting the biological relevance for contacts detected in both datasets (**Figure 3.11**).

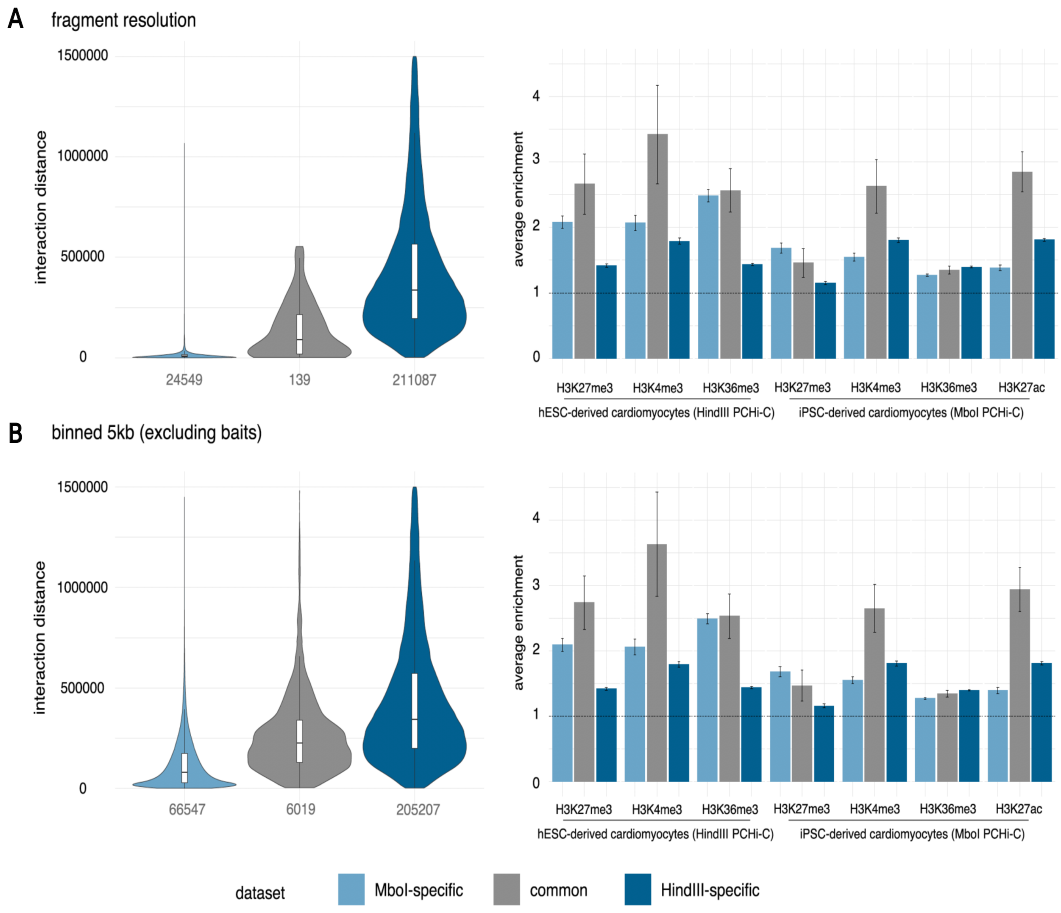


Figure 3.11: **Comparative analysis of Capture Hi-C data generated with *MboI* and a *HindIII* restriction enzyme.** Three *MboI* Promoter CHi-C (PCHi-C) replicates for iPSC-derived cardiomyocytes (iPSC CMs33) were processed by CHiCAGO either at the restriction fragment level or by grouping adjacent restriction fragments in 5kb bins. Three *HindIII* PCHi-C replicates for hESC-derived cardiomyocytes (hESC CMs34) were processed using standard CHiCAGO settings. Interactions with a CHiCAGO score $\geq$5 were considered shared when the middle of the significantly interacting fragments in the *MboI* PCHi-C data fell within the respective interacting fragments in the *HindIII* dataset. Comparison between *MboI* and *HindIII* PCHi-C datasets in using *MboI* non-binned data, panel **A** and between *MboI* binned PCHi-C data, panel **B**. Violin plots show the distance distribution of significant interactions shared between *MboI* and *HindIII* (grey), *MboI*-specific (light blue) and *HindIII*-specific (dark blue). The number of interactions in each group is specified on the x-axis. The difference in the distance range of interactions observed between *MboI* and *HindIII* is mainly driven by the average restriction fragment size given by two enzymes. As shown, the binning strategy allows to partially rescue longe range interactions in *MboI*-processed data. The barplots show the enrichment for regulatory histone marks (as a ratio between observed and expected) in each group of interactions. Figure from Freire-Pritchett, P. and Ray-Jones, H. et al, 2021 [433].

Choosing a four-cutter enzyme can also affect how CHiCAGO calculates the background when calling signification interactions. CHiCAGO takes into account two main sources of noise for the estimation of the background: the Brownian component and the technical component. It then combines the two components to estimate the background and it computes normalization (or scaling factors) for each interaction pair detected. If the background is estimated inaccurately and sparsity is not accounted for appropriately it can impact the calculation of the scaling factors and, consequently, affect the final p-values assigned to each interaction pair. Therefore, I set out to establish the way to determine the correct parameters for a proper CHiCAGO background estimation when using a four-cutter enzyme.

An incorrect estimation of background can be visually observed in the CHiCAGO interaction profiles, as shown in **Figure 3.12**, top left panel. A rapid decay of the brownian component towards zero is usually a good indication of inaccurate background which leads to calling significant interaction pairs with very low read counts. To define the correct parameters for appropriate background estimation I assessed data sparsity by computing the proportion of missing interactions (i.e. interactions showing zero counts) per distance bins across baits, for a customized range of values for `maxLBrownEst`.

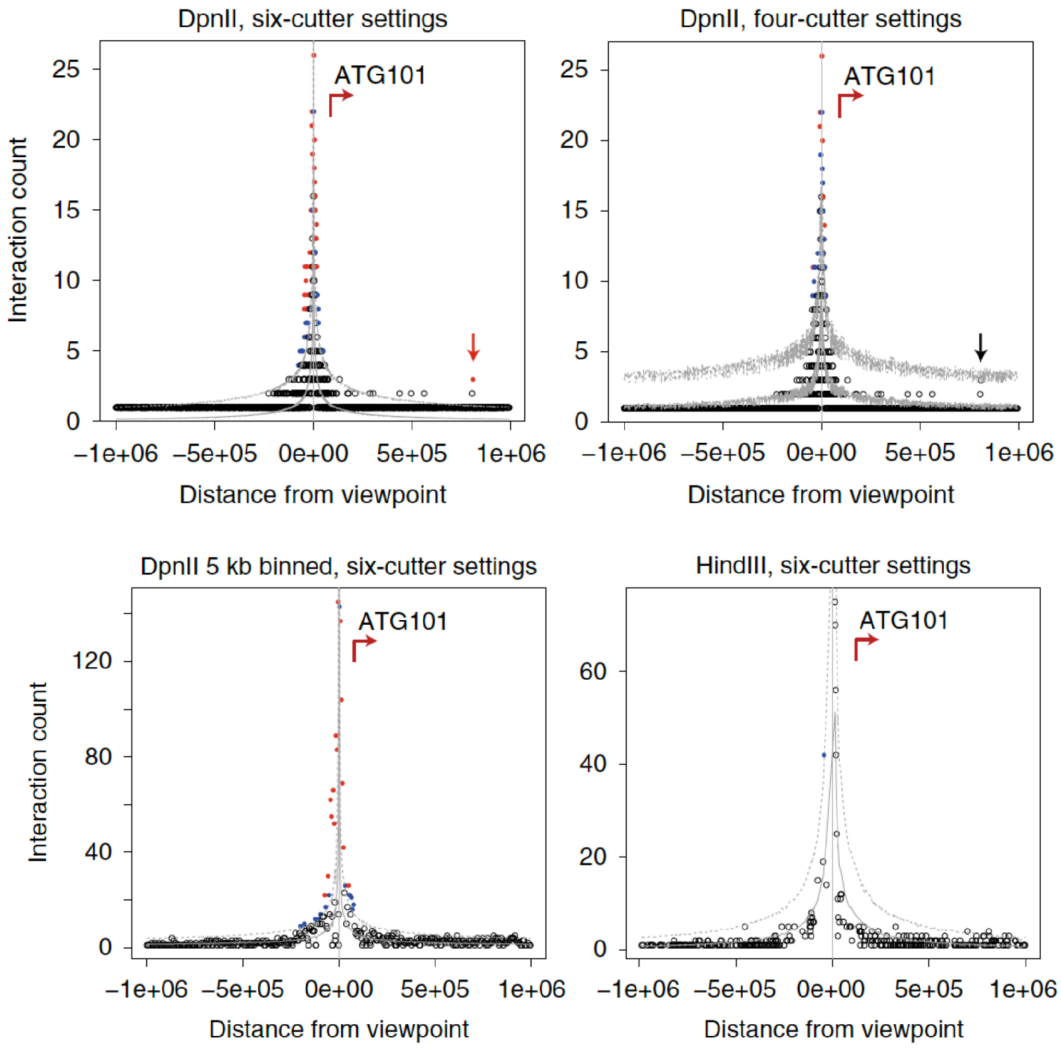


Figure 3.12: **Visualizing the incorrect estimation of CHiCAGO background model.** *DpnII* PCHi-C data in iPSCs-derived cardiomyocytes [285] were analyzed using six-cutter CHiCAGO parameters or using the `maxLBrownEst` and `binsize` parameters suggested for four-cutter enzymes. *DpnII* data binned into 5kb bins was also analyzed using CHiCAGO default six-cutter parameters along with *HindIII* PCHi-C data. As an example, the figure shows the interaction profile of the *ATG101* promoter for each of the four processed datasets: when *DpnII* PCHi-C data are analyzed using the default CHiCAGO parameters (top let panel) a spurious low-count interaction is detected as significant (red arrow), which does not appear in the other three data/parameter combinations. Figure from Freire-Pritchett, P. and Ray-Jones, H. et al, 2021 [433].

When applying the default settings of CHiCAGO to *DpnII*-derived CHi-C data I observed a rapid increase in the proportion of missing interactions as distance increases. On the contrary, this it is not the case for *HindIII*-derived CHi-C data processed with CHiCAGO's default parameters (**Figure 3.13**).

I then changed `maxLBrownEst` and `binsize` to 75kb and 1.5kb respectively. `maxLBrown Est` was empirically determined to be sufficiently large to observe a typical brownian distance decay, but at the same time small enough to avoid baits to become overly sparse across the given distance range. For `binsize` I chose a value that is at least five times longer than the average restriction fragment size. The rationale being that the bins need to be large enough to estimate the average read count robustly, but also not so large that the count of individual interaction pairs within the bins vary too much in read coverage. **Figure 3.13**, shows the estimation of missing interactions at given distance bins when processing *DpnII*-derived CHi-C data using `maxLBrownEst` set at 75kb and `binsize` set at 1.5Kb. With these parameters, the data displayed a more gradual decay, similar to what observed for *HindIII*-derived CHi-C data analyzed with default CHiCAGO parameters (**Figure 3.13**). A comparable behaviour can be observed for 5kb-binned *DpnII* data (**Figure 3.13**).

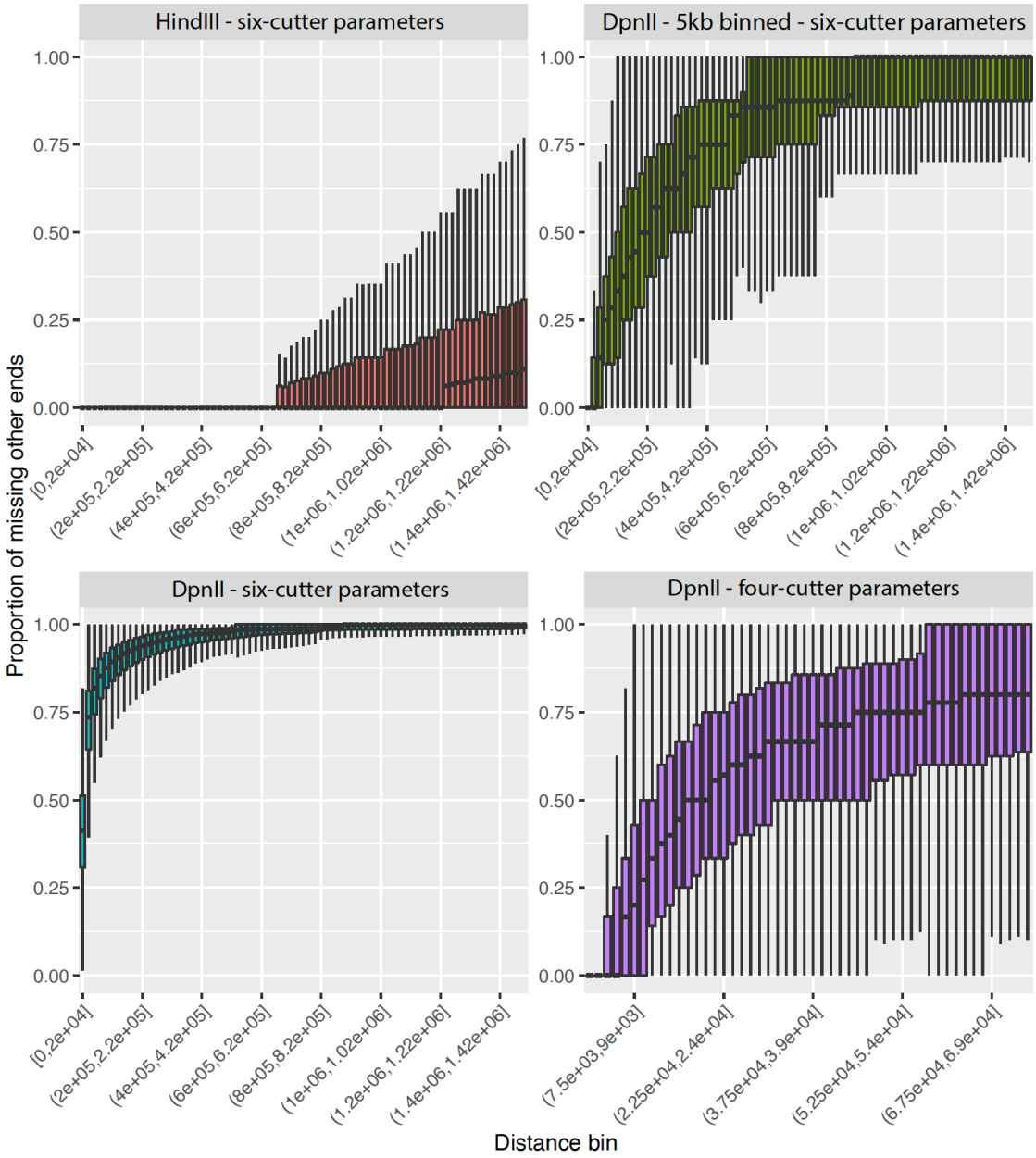


Figure 3.13: **Estimation of missing interactions for CHiCAGO background model.** Boxplots indicating data sparsity at different distance bins for estimating the Brownian background component (the distance range plotted corresponds to `maxLBrownEst`, and the size of each bin is set to `binsize`). Sparsity was defined as the proportion of all other-end fragments within the bin with a count of zero. For *DpnII* with four-cutter parameters, 5kb binned *DpnII* and *HindIII* with default parameters, sparsity increases gradually, while for *DpnII* CHi-C data analyzed with six-cutter settings sparsity rapidly increases, with almost all other ends of interactions with a count of zero for most baits. The boxplots were generated using the `plotBackgroundSparsity.R` script provided as part of `chicagoTools`.

Here, I devised a way to fine tune CHiCAGO parameters in order to analyze and detect

significant interactions for *DpnII*-derived CHi-C data, with comparable sensitivity and specificity to the analysis of data generated with a six-cutter enzyme when using default CHiCAGO parameters. I established a way to assess the correct estimation of background, hence the appropriate estimation of the normalizing factors and p-value, in order to avoid low read count interactions being called as significant (**Figure 3.12**). In addition, I devised a binning strategy to help mitigate the inherent sparsity given by four-cutter data and to partially rescue longer-range interactions detection [433].

3.2.5.2 CHiCAGO score cutoff to call "significant" interactions

By default, CHiCAGO calls interactions "significant" if they pass a defined score of 5 (based on previously described analysis [432]). I next sought to define a possible strategy to further tune the score threshold given by CHiCAGO to identify significant interactions. The strategy I adopted to define the choice of the score cutoff for a given experimental setting is based on the identification of the right balance between the enrichment for H3K4me1 at other ends, a biologically relevant chromatin feature for the recognition of enhancers, and the recall of H3K4me1 peaks, as shown in **Figure 3.14**.

This strategy provides a method for determining *ad hoc* cutoffs when using customized parameters and based on the specific research question.

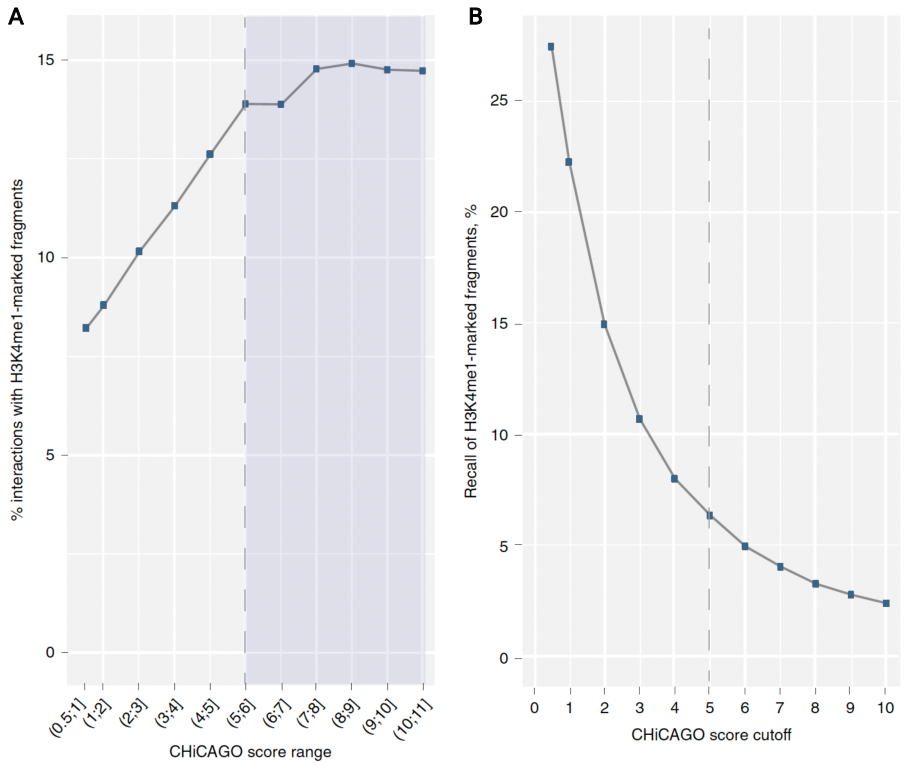


Figure 3.14: **Tuning the CHiCAGO score cutoff by balancing recall and enrichment of regulatory chromatin features. A.** Percentage of interactions with H3K4me1 marked fragments (y-axis) within a given CHiCAGO score range (x-axis), computed for *HindIII*-generated Capture-HiC data. Highlighted in blue the CHiCAGO scores range at which the enrichment of interactions with H3K4me1-marked fragments starts to plateau. **B.** Recall of H3K4me1 marked fragments (y-axis) at the increasingly stringent CHiCAGO score cutoffs (x-axis). The gray dashed line highlights default CHiCAGO score cutoff of 5. Figure from Freire-Pritchett, P. and Ray-Jones, H. et al, 2021 [433].

3.3 DISCUSSION

This chapter presents the steps for the further refinement of critical steps of the miniaturized CHi-C protocol starting from a small cell number as input, previously developed in our lab by Malysheva V. [452], and the adjustments of current analysis pipelines, HiCUP and CHiCAGO, for processing *DpnII*-derived CHi-C data.

I first focused on the optimization of crucial aspects of the Hi-C stage of the protocol. Here, I showed that our TR-Tn5 performs equally well to the commercially available TDE and I defined optimal experimental conditions to achieve high quality Hi-C and CHi-C libraries, introducing specific adjustments when processing hESCs. In parallel, I focused on the calibration of current pipelines, HiCUP and CHiCAGO, routinely used in the lab for the downstream analysis of Hi-C and CHi-C data and I presented a refined approach for the analysis of four-cutter derived Hi-C and CHi-C, such as *DpnII*-generated libraries.

In summary, the results confirmed that our calibration of the HiCUP pipeline resulted in the rescue of valid sequencing pairs after the alignment and the processing of *DpnII*-based libraries, with consequent gain of informative sequencing reads. Moreover, this chapter presented the fine tuning of key parameters of the CHiCAGO pipeline for the appropriate identification of significant interactions (now a publicly available protocol [433]).

I explored alternative approaches to implement the conventional data analysis of capture Hi-C data in order to, ultimately, make its interpretation more comprehensive and adaptable to different datasets (e.g. data generated by the use of restriction enzymes other than the commonly used *HindIII* and *DpnII*) and/or specific biological questions.

3.3.1 Reaching the balance between enrichment and yield to achieve the necessary sequencing depth

Alongside *DpnII* enzyme to increase CHi-C resolution, the use of Tn5-transposase represents a great improvement of the capture Hi-C protocol.

The increased use of next-generation sequencing (NGS) technologies sparked the necessity to make the processing of the samples easier and quicker. With its "cut and paste" system, transposition mediated by Tn5-transposase it has been implemented in many NGS-based assays, making sample processing more efficient and simple, considerably shortening library construction preparation and enabling the realization of large-scale se-

quencing projects. Indeed, the use of Tn5-mediated transposition is amenable to adjustments or implementations in a variety of NGS protocols: for example it is now regularly employed in DNA accessibility assays (ATAC-seq) [454]; it is used in the recently developed chromatin profiling assay, CUT&TAG (Cleavage Under Target and Tagmentation) [425], or in the most recently described alternative 3D conformation capture protocol, BAT Hi-C (Bridge linker-Alul-Tn5 Hi-C) [455].

However, alongside its clear advantages, the use of Tn5-trasposase can present several limitations, amongst which its elevated costs, which can make large-scale sequencing projects inaccessible. In this chapter I demonstrated that I can overcome this limitation with the production of a TR-Tn5 which performs at the same level as the commercially available one, generating high-quality Hi-C and CHi-C libraries ready for sequencing.

In addition to costs, other factors can represent a challenge when adapting protocols to the use of Tn5-transposase. In the particular case of our CHi-C protocol, the introduction of Tn5-transposase for the tagmentation of ligated DNA, to substitute for DNA sonication ad size selection, led to a final Hi-C library with a less controlled range of fragments size. In addition, Tn5-transposase mediates a stochastic introduction of adapters which can lead to the generation of a final Hi-C library that contains smaller fragments that may not be representative of "true" contacts. Ultimately, this affects the final Hi-C quality, reducing the percentage of valid pairs, thus leading to a decrease in available sequencing reads and, effectively, decreasing the final sequence coverage. Especially in the case of CHi-C, low sequencing depth raises sparsity issues that impacts on the ability to robustly identify significant interactions in downstream analysis and it eventually makes CHi-C experiments inaccessible due to the increasing need of additional sequencing to reach the number of reads necessary.

Interestingly, I observed that different cell types can perform differently when processed for Hi-C library preparation adopting Tn5-mediated tagmentation. This has been the case for hESCs, for which standard parameters of the protocol did not represent the optimal conditions to generate good quality Hi-C and CHi-C libraries. Usually, the size distribution of the fragments observed in both naïve and primed hESCs appears to be shifted towards smaller sizes (with naïve hESCs showing the larger shift, **Figure 3.6, A**). It is known that the chromatin in these cells tends to be in a more accessible state. Presumably, this could lead to the generation of shorter fragments following the *DpnII* digestion,

increasing the possibility of generating fragments that are not representative of "true" interactions. Although in the protocol presented in this chapter it does not directly affect Tn5-mediated tagmentation, DNA accessibility might affect the initial restriction enzyme digestion, leading to the generation of shorter fragments. As a consequence, the chances for Tn5 to stocastically insert adaptors within these shorter fragments increases, resulting in a greater proportion of products non representative of "true" interactions in the final Hi-C and CHi-C libraries. Moreover, aspects like the ratio between the amount of Tn5 enzyme and DNA and/or the presence of impurity, such as glycerol, have to be precisely controlled to avoid a non-uniform insertion of adapters mediated by Tn5 [456, 457, 458]. Therefore, different aspects of the Tn5-mediated tagmentation reaction can raise issues of sensitivity in the final library. In our specific case, for example, sub-optimal conditions seemed to affect the percentage of valid pairs of the final library, favoring the presence of reads mapping within the same restriction fragment. While this could be due to an incorrect insertion of adapters by Tn5, it does not provide a comprehensive explanation for the lower percentage of valid pairs I observed in some cases, which is likely to be affected by additional, and perhaps complementary, factors at different stages of the protocol.

Nevertheless, with the adjustments presented in this chapter, I defined the optimal protocol conditions to robustly generate good quality Hi-C and CHi-C libraries with the use of tagmentation ready Tn5-transposase.

CHi-C presents the additional challenge of obtaining a high degree of enrichment for the regions of interest. As mentioned earlier, this can be crucial in determining the final number of available reads, hence the final sequencing depth necessary for downstream analysis. Therefore, obtaining a high enough capture efficiency is key to make large-scale sequencing projects feasible.

The advantage of using a four-cutter enzyme like *DpnII* as opposed to a six-cutter enzyme (such as *HindIII*) in our approach is the increased resolution ($\sim$ 15-fold). This, however, requires a considerably higher sequencing coverage. Maximizing the percentage of valid pairs and capture efficiency minimizes the loss of usable reads after Hi-C or CHi-C sequencing data are aligned and processed. Likewise, preserving the yield is vital to be able to obtain the number of reads necessary: only a limited number of reads can be obtained from a given amount of library. Therefore, it was crucial to identify the conditions, described in this chapter, to reach the right balance between enrichment and final yield when generating CHi-C libraries, both fundamental to obtain the necessary final

sequencing depth.

3.3.2 Fine tuning analysis of four-cutter restriction enzyme derived CHi-C data

CHi-C data allows to identify interactions of regions of interest where each captured fragment can interact with one or many regions of the genome. This gives CHi-C data an asymmetric nature that confers to it unique statistical properties. The lab previously developed CHiCAGO, a pipeline for the analysis of CHi-C data and the detection of significant chromosomal interactions. However, as previously mentioned, CHiCAGO has been initially designed and fine tuned for CHi-C data derived from a six-cutter enzyme (i.e. *HindIII*) [432].

It is now clear that, depending on the restriction enzyme, the properties of the resulting data can change, requiring *ad hoc* parameters in order for CHiCAGO to correctly estimate the background noise and, consequently, call true significant interactions.

As mentioned earlier, a different resolution is one of the most obvious differences arising from CHi-C data processed with a four-cutter as opposed to a six-cutter enzyme. A four-cutter enzyme like *DpnII* cuts the human genome (that consists of 3 x 10^9 base pairs) $>$ 11,000,000 as opposed to >730,000 like in the case of a six-cutter enzyme like *HindIII*. This can translate into greater sparsity observed in *DpnII*-derived CHi-C. In this chapter I provided examples of how this can influence the ability of CHiCAGO to identify true significant interactions and a possible strategy to mitigate such sparsity by creating bins that group adjacent restriction fragments, hence increasing the number of reads per viewpoint. I showed that binning also represents a valid strategy to rescue longer-range distance interactions that are not detected in four-cutter processed CHi-C data. This is likely due to the drastic difference in fragment size produced by the different restriction enzymes, on average $\sim$400bp fragments for *DpnII* compared to an average of $\sim$5kb in the case of *HindIII*. As previously shown in this chapter, I can observe a great difference in distance range of interactions between four-cutter and six-cutter derived CHi-C data, with a very small overlap between the interactions detected in the two different datasets. However, I showed that grouping adjacent restriction fragments in four-cutter derived CHi-C data has to potential to partially compensate for the "loss" of long-range interactions. Interestingly, the same was observed in an independent study from Su et al, 2021 [459, 433].

Therefore, a valid strategy for the analysis of CHi-C generated using a four-cutter en-

zyme could combine the two approaches, at different resolution levels: firstly, the analysis of high resolution CHi-C data, that are more sensitive to the detection of more proximal chromatin interactions, which can prove valuable, for example, in GWAS studies where the focus is on the identification of regulatory regions enriched in disease-associated variants. Then, the complementary approach of *in silico* binning of the adjacent restriction fragments to rescue, in part, longer-range interactions.

In addition, I probed a different interpretation for the CHiCAGO score and the definition of its threshold. By default, CHiCAGO considers an interaction significant if it scores $\geq$5. But CHiCAGO scores represent weighted p-values, based on previously described analysis [432], and they serve mostly as ranking measures. Therefore, the choice to apply a specific score cutoff to determine whether or not to consider an interaction significant is liable to subjectivity. While CHiCAGO score still represents a valuable tool for the prioritization of interactions of interest, I propose that it should not be used in a dualistic manner to determine whether a contact is "present" or "absent". Instead, I suggest that one possibility would be to consider such score in a quantitative manner instead, through approaches such as clustering and/or regression modeling. For example, an alternative approach is described in a recent work from Disney-Hogg et al., 2020 [460] where the authors recommend to define a score cutoff based on the reproducibility of interaction calls between replicates.

Overall, the fine tuning process for the downstream analysis described in this chapter (and in more details in [433]), it is not limited in its application to data generated by the use of a four-cutter or a six-cutter enzyme alone. For example, the presented binning strategy could be extended to an incremental binning approach, whereby bins of different sizes are applied to assess the overlap between datasets and for the analysis of interactions within different distance ranges. Ideally, the estimation of CHiCAGO background, with consequent adjustment of p-value weighing and a more *ad hoc* choice of a score threshold can be applied to CHi-C data generated with any different restriction enzymes: for example, it could prove valuable in the case of Micro- C and Capture Micro-C data, characterized by a even higher resolution than the one given by *DpnII*, hence giving the data different features to be taken into account in the downstream data analysis.

3.3.3 Conclusion

In conclusion, this chapter provides the specification of the preparatory work that has been necessary in order to establish a robust protocol for the generation of good quality Hi-C and CHi-C, with adjustments specific to hESCs, on both the experimental and computational level.

The protocol and its refined steps presented in this chapter enabled the generation and the analysis of the data to explore PEs regulatory dynamics upon the naïve-to-primed transition and shed light on their potential functional role in pluripotency and cell-fate determination, which will be the focus of the next chapter.

4 The emergence of poised enhancers (PEs) upon naïve-to-primed transition of human embryonic stem cells (hESCs)

4.1 INTRODUCTION

Originally described in ESCs, PEs represent a specific class of CREs defined by the joint presence of the 'active' H3K4me1 and the 'repressive' H3K27me3. It has been hypothesized that PEs may be part of a fine gene regulatory mechanisms to promote timely activation of genes for proper differentiation to occur, mediating a rapid switch between a repressive and an active state

The presence of H3K27me3 at PEs suggests the recruitment of Polycomb-group proteins (PcGs) to these sites. Besides their established role in gene repression during development, PcGs play an important role in mediating long-range interactions in embryonic stem cells (ESCs). Moreover, it is known that PcGs undergo significant reorganization upon the naïve-to-primed transition in human ESCs (hESCs), during which a rearrangement of both H3K27me3 chromatin patterns and Polycomb-mediated interactions occurs. Specifically, primed hESCs display higher degree of interconnection between Polycomb-interacting regions compared to their naïve counterpart [278]. Therefore, it can be speculated that, likewise, PE-mediated interactions might undergo extensive rewiring when cells transition from the naïve to the primed state of pluripotency.

Through experimental and computational approaches, here I investigate the emergence of the poised state of enhancers and their contacts upon the naïve-to-primed transition in hESCs through a Poised Enhancer Capture Hi-C (PECHi-C) approach and chromatin profiling Cut&Tag assays. Furthermore, this chapter presents the establishment of an inducible CRISPRa (iCRISPRa) system in iPSCs [423] for the perturbation of selected

candidate PEs, with the aim to shed light on the potential role of the poised state of enhancers in pluripotency and cell-fate decision.

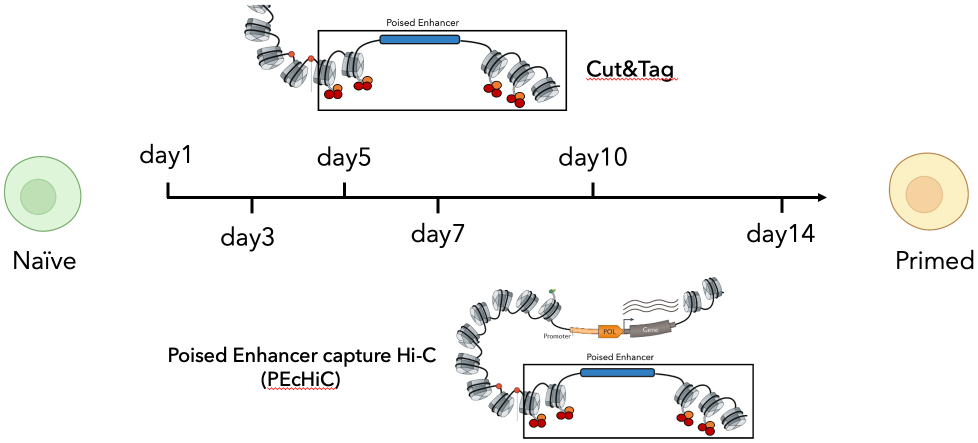


Figure 4.1: **The experimental approach used to profile the emergence of PEs upon naïve-to-primed transition in hESCs.** The emergence of the poised state of enhancers and their interaction network was profiled through a time course as hESCs transitioned between the naïve state of pluripotency and the primed state. Eight different timepoints were profiled using PECHi-C and Cut&Tag assays: naïve (hNES1), day 1 (hNES1), day 3 (hNES1), day 5 (hNES1), day 7 (hNES1), day 10 (hNES1), day 14 (hNES1) and primed (H9) hESCs.

4.2 RESULTS

4.2.1 Devising a Poised Enhancer Capture Hi-C (PECHi-C) system

In order to build a comprehensive catalogue of PEs, I identified putative PE regions based on their bivalent chromatin signature, making use of publicly available H3K4me1 and H3K27me3 ChIP-seq datasets (ENCODE, [461]) and Cut&Run datasets (Rugg-Gunn's lab, unpublished) generated in primed H9 hESCs. The overlap of H3K27me3 and H3K4me1 signals resulted in the identification of 54,363 regions enriched for both histone marks.

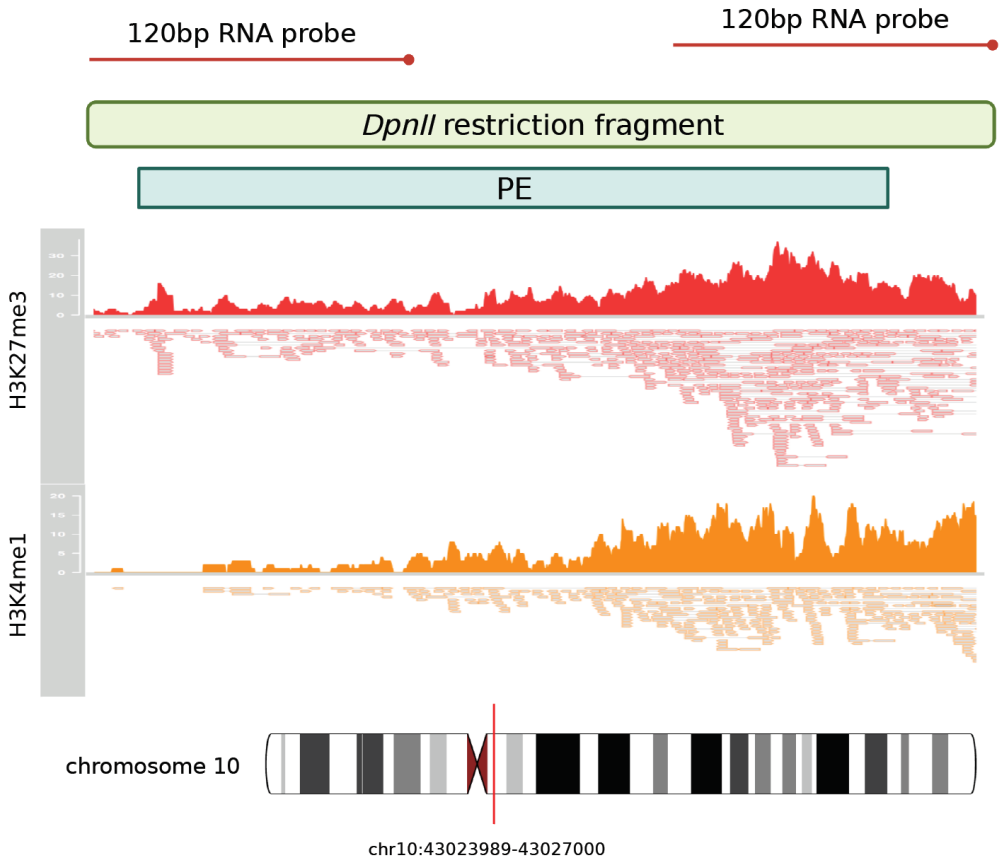


Figure 4.2: **Viewpoint of a PE region included in the PECHi-C capture system approach.** IGV (Integrative Genomics Viewer) [462] viewpoint of a typical PE region (dark green bar) included in the PECHi-C capture system, showing the enrichment of H3K27me3 and H3K4me1. RNA biotinylated probes (in red) were designed to be complementary to *DpnII* restriction fragments (light green bar) overlapping all identified PE regions. PECHi-C final capture system includes ≥60,000 RNA probes, with an average of two probes per *DpnII* restriction fragment overlapping the enhancer region of interest.

After filtering out low-confidence regions (i.e. ChIP-seq enrichment peaks identified in ≤2 of the data sets analyzed) and regions within a 1kb window from annotated gene TSSs [436] in order to exclude main gene promoters, **49,310** putative PEs were retained and used to design the pool of complementary biotinylated RNA probes for hybridization with our regions of interest (henceforth referred to as the "capture system"). Specifically, 120bp RNA probes labelled with biotin have been designed to capture *DpnII* digestion fragments overlapping with the identified PE regions: the final capture system included a total of **60,764**, with an average of two RNA-biotinylated probes hybridizing each restriction fragment (**Figure 4.2**).

This PECHi-C approach, devised to specifically capture PE regions, offered the ad-

vantage of increased sensitivity with respect to contacts that directly involve PEs, at least at one end, as opposed to the more commonly used Promoter Capture Hi-C (PCHi-C) approach in which RNA-biotinylated probes are designed to be complementary to all annotated promoters [275].

4.2.2 Different poised enhancer interaction dynamics upon the naïve-to-primed transition

To profile the emergence of PE-mediated contacts, I performed PECHi-C experiments upon the naïve-to-primed transition in hESCs (**Figure 4.1**) and in hESCs differentiated to definitive endoderm and neuro-ectoderm.

Initial quality control analyses of the generated PECHi-C data showed a percentage of valid pairs ranging between 50% and 80% and a capture efficiency between 16% and 50% of the final CHi-C libraries, achieving a fold enrichment for our regions of interest between 20 and 60 fold (**Table 4.1**). Pearson's correlation analysis was also performed to assess the robustness of biological replicates (**Supplementary Figure A.3, Appendix A**, with all replicates displaying an $R^2 \geq 0.5$ [432].

timepoint	Total number of reads (PE)	% of valid pairs	Capture Efficiency
Replicate 1			
naïve	143,854,542	53%	20.16%
day1	152,567,161	60%	17.66%
day3	105,849,412	58%	35%
day5	110,834,728	56%	23.67%
day7	68,233,873	59%	45.48%
day10	106,331,789	39%	19.92%
day14	45,850,977	50%	37.78%
primed	50,225,481	72%	42.56%
Neuro Ectoderm	155,603,514	73%	19.46%
Definitive Endoderm	228,798,705	76%	16.6%
Replicate 2			
naïve	114,360,576	52%	24.42%
day1	65,786,309	55.5%	19.69%
day3	45,022,769	60%	33.37%
day5	125,157,239	55%	26.14%
day7	71,044,786	49%	25%
day10	84,874,382	48.2%	47.01%
day14	80,597,754	56%	27.74%
primed	30,612,355	65%	34.71%
Neuro Ectoderm	149,143,031	72%	20.66%
Definitive Endoderm	159,630,501	64%	18.1%

Table 4.1: **QC summary of PECHi-C libraries.** The table summarizes the total number of paired end (PE) sequencing reads obtained for each sample, the final PECHi-C library quality, calculated as percentage of final valid pairs and the fold enrichment of the final PECHi-C library for interactions involving PE regions of interest, represented as capture efficiency (percentage of reads exclusively mapping to the regions of interest).

I identified **182,303** interactions with CHiCAGO score ≥5 in at least one of the timepoints of the transition and I first asked if a potential trend in the acquisition of PE interactions over the time course could be identified as cells transitioned into the primed state. Initial PCA analysis of the individual timepoints revealed a trajectory that generally followed the progression of the naïve-to-primed transition along dimension 1 (**Figure 4.3, A**). However, dimension 2 was not entirely independent of dimension 1, showing the so called *"horse shoe"* effect that suggested a non-linear relationship between the different timepoints and it was not able to identify one predominant dimension to describe PECHi-C data (as also suggested by the screeplot in **Figure 4.3, C**). While PCA showed a distinction between differentiated cells and hESCs, as well as between primed hESCs and the earlier stages of the transition, the earlier timepoints however clustered together with no clear dependency on time (**Figure 4.3, B**). This was in contrast with the results shown by differential gene expression PCA analysis, which instead showed a clear dependency on time already at the very early stages of the transition (**Supplementary Figure A.4, Appendix A**). Indeed, the PECHi-C approach was designed to specifically capture contacts directly involving PEs, which in most cases do not affect the expression of their target genes until later during differentiation, despite pre-establishing their 3D contacts in hESCs [233, 264]. Therefore, the changes observed in PE-mediated contacts in hESCs are not entirely reflected in the difference of gene expression observed.

I then applied a hierarchical clustering approach for the partitioning of the single timepoints. Similarly to PCA analysis, hierarchical clustering highlighted two main branches formed by differentiated cells and hESCs, but it generated a shallow dendrogram for the earlier timepoints, suggesting that cells at earlier stages of the transition possibly retained many features of the naïve state (**Figure 4.3, D**).

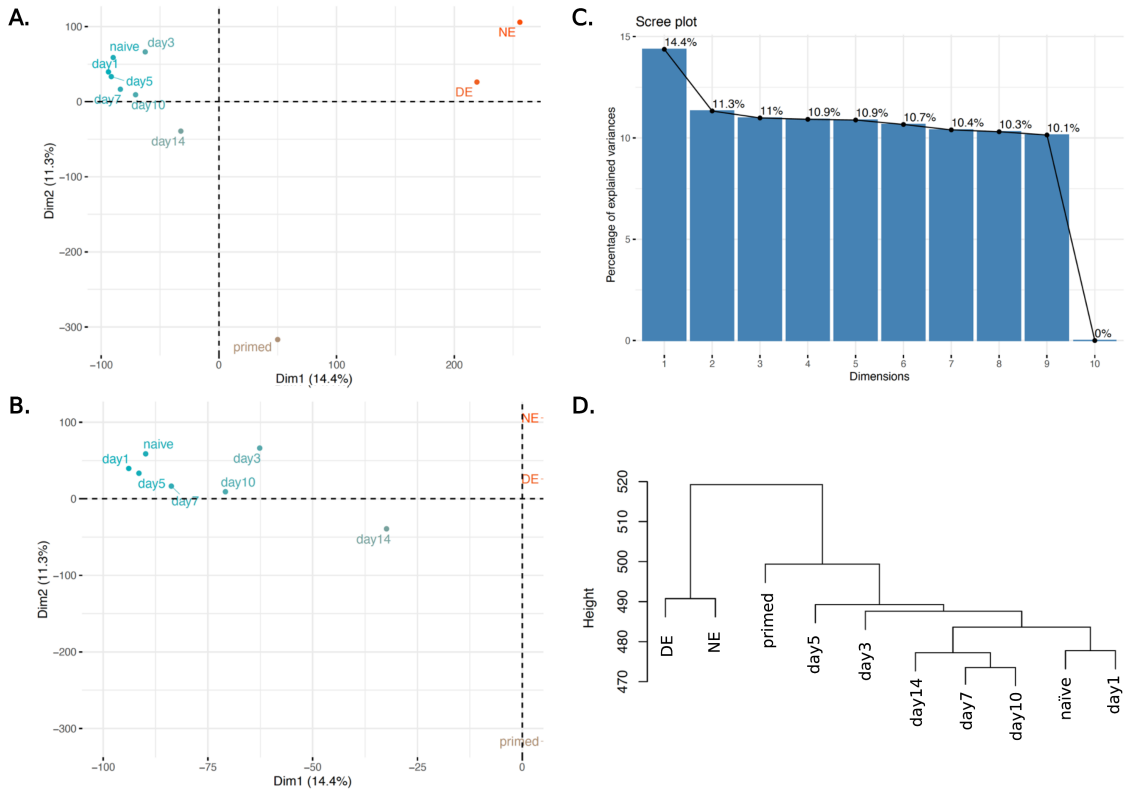


Figure 4.3: **PCA and hierarchical clustering analysis of individual timepoints of the naïve-to-primed transition.** Dimensionality reduction and clustering approaches applied to PECHi-C data generated for hNES1 hESCs during the time course of the naïve-to-primed (primed H9 hESCs) transition and differentiated hNES1 (DE and NE). Interactions with a CHiCAGO score ≥5 in at least one of the timepoint were considered for the analysis. **A.** PCA analysis of individual timepoints of the naïve-to-primed transition. Dimension 1 (x-axis) explains 14.4% of the variability and suggests a trajectory that follows the progression of hESCs between the two states of pluripotency, while Dimension 2 represents 11.3% of the variablity. **B.** A zoomed in view of the the left top corner of panel A into the earlier timepoints of the transition. **C.** Scree plot showing the percentage of variability explained by 10 dimensions identified through the PCA analysis. PCA analysis did not identify one main dimension to describe PECHi-C data. **D.** Hierarchical clustering dendrogram of single timepoints of the naïve-to-primed transition showing two main branches: left branch formed by the differentiated hNES1 (i.e. NE and DE) and the right branch formed by the earlier timepoints of the hNES1 naïve-to-primed (primed H9 hESCs) transition.

In addition, the hierarchical clustering analysis also generated multiple clusters characterized by pronounced interaction signals at only one timepoint, which could be driven by data sparsity. To verify and potentially mitigate such sparsity, I attempted to impute missing counts using expected counts at a given interaction distance. Briefly, after computationally compiling a list of candidate PE-mediated contacts within the distance-range

window of the missing data points (i.e. 1,000bp - 10,000bp, **Figure 4.4, A**), I then assigned to the implemented PE-mediated contacts the number of reads (N) imputed using the CHiCAGO distance function of contacts with matching distance-range (see **Methods**). However, PCA and hierachical clustering analyses of the imputed data showed that this approach likely resulted in over-fitting, masking a possible inter-dependency between PE contacts and time (**Figure 4.4, B** and **C**). This effect could also be appreciated when partitioning the data based on the arcsine transformed CHiCAGO scores of contacts, instead of single timepoints of the transition. Indeed, k-means clustering of the imputed contacts confirmed that the imputation of the PECHi-C data removed any possible temporal dependency (**Figure 4.4, D**).

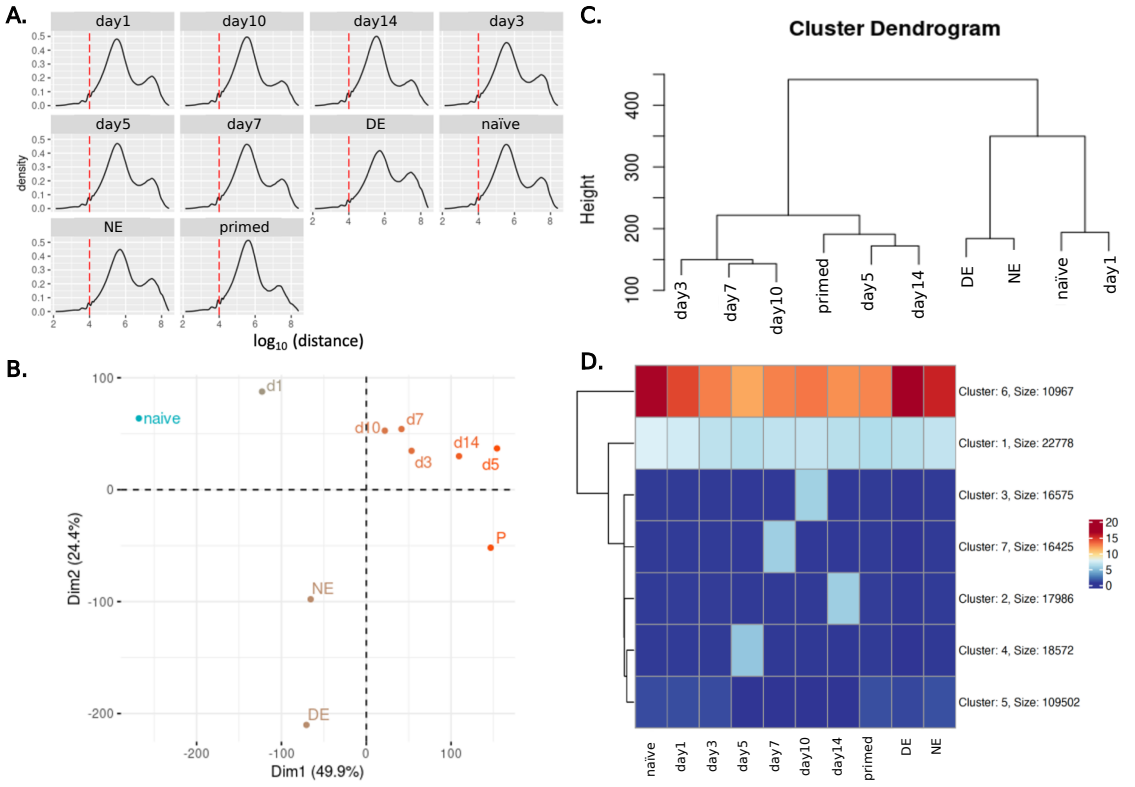


Figure 4.4: **Analysis of imputed PECHi-C data using PCA, hierarchical clustering and k-means clustering.** Dimensionality reduction and clustering approaches applied to imputed PECHi-C data generated for hNES1 during the time course of the naïve-to-primed (primed H9 hESCs) transition and differentiated hNES1 (DE and NE). Interactions with a CHiCAGO score $\geq$5 in at least one of the timepoints were considered for the analysis. **A.** Density plots showing the distribution of interactions distance range (x-axis, showed as $\log_{10}$ of distance) of PECHi-C data upon hNES1 naïve-to-primed (primed H9 hESCs) transition and hNES1 differentiated cells (DE and NE). The red-dashed line marks the cutoff chosen to compile short-range candidate PE-mediated interactions for the imputation strategy. **B.** and **C.** PCA and hierarchical clustering analysis of the single timepoints using imputed PECHi-C data. **D.** k-means clustering partitioning of imputed PECHi-C contacts. The resulting heatmap shows that the imputation approach likely resulted in the over-fitting of the data, removing any possible temporal dependency.

I then partitioned non-imputed PECHi-C contacts in order to probe the possible link between time and PE connectivity and specifically identify contacts with a temporal dependency. However, k-means clustering did not reveal a clear dependency between contacts and time (**Figure 4.5**). Nevertheless, this did not preclude the possibility that subclasses of interactions showing temporal dependence existed in the data, since clustering is not geared to specifically look for patterns of this kind.

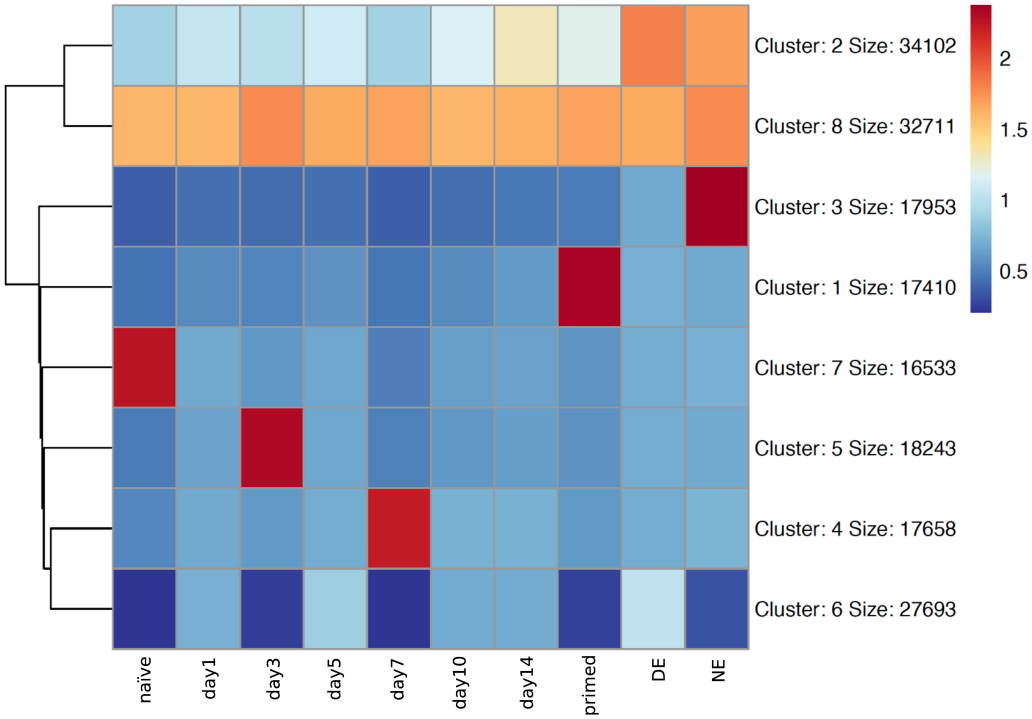


Figure 4.5: **K-means clustering analysis of non-imputed PECHi-C data.** K-means clustering approach applied to non-imputed PECHi-C data generated for hNES1 during the time course of the naïve-to-primed (primed H9 hESCs) transition and differentiated hNES1 (DE and NE). Interactions with a CHiCAGO score $\geq$5 in at least one of the timepoint were considered for the analysis.

Therefore, I employed a more targeted approach to identify contacts that were acquired, lost or retained over the course of the naïve-to-primed transition. For this, I computed Spearman's rank order correlation to assess the association between PE-mediated contacts and time and to determine the direction of such association, mainly focusing on PE-mediated interactions with annotated TSSs and interactions between PEs.

Using this approach, I defined three interaction classes: the **UP** class ($\rho \geq$0.4, n = 2,433), **DOWN** ($\rho \leq$-0.4, n = 1,357) class and **CONSTANT** class (-0.2$\leq \rho \geq$0.2, n = 1,817).

Notably, it could be observed that contacts in the UP class showed a significant increase in CHiCAGO scores on **day 3** of the transition (**Figure 4.6**. Box plot on the right corner of the top panel shows a significant shift in the median of the CHiCAGO score distributions between day 1 and day 3 of the naïve-to-primed transition, Wilcoxon test p-value < 0.0001. Such increase then appears to become more gradual as cells progress from day 3 into the primed state, as can be observed in the violin plots in blue, UP class), highlighting this timepoint as potentially critical for the emergence of PE-mediated contacts. The DOWN class described an opposite trend, whereby interactions were gradually lost as cells progressed into the transition, while the CONSTANT class included a group of interactions that remained stable over the time course, perhaps suggesting that, in some cases, PE-mediated interactions were "pre-set" in naïve cells and maintained as cells transitioned between the two states of pluripotency (**Figure 4.6**).

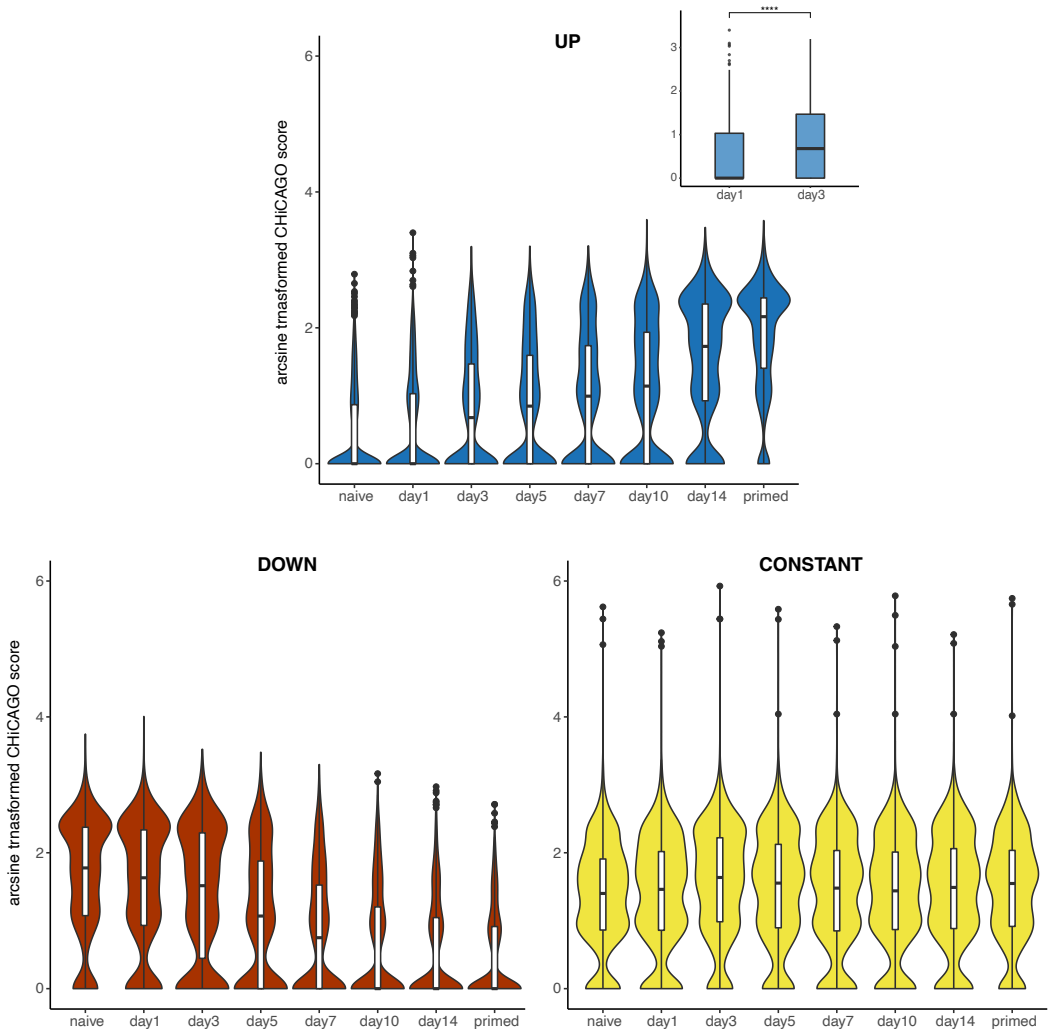


Figure 4.6: **Dynamics of the emerging of PE-mediated contacts upon the naïve-to-primed transition.** Violin plots showing the distribution of CHiCAGO scores (y-axis, represented as arcsine trasformed CHiCAGO scores) for each timepoint of the naïve-to-primed transition (x-axis) for interactions in the UP interactions class (blue, Kruskal-Wallis, p-value $\leq$ 2e-16), the CONSTANT class (yellow) and the DOWN class (red, Kruskal-Wallis, p-value $\leq$ 2e-16). Box plot (top right panel) shows in more detail the significance difference of CHiCAGO score distributions between day1 and day3 of the transition in the UP class (Wilcoxon test, p-value $<$ 0.0001, ****). PECHi-C for hESCs that showed a CHiCAGO score of $\geq$5 in at least one of the timepoints were included. Differentiated hNES1 (NE and DE) were excluded from this analysis.

Overall, these analyses identified a considerable number of PE-mediated contacts that showed the temporal patterns described above. In particular, Spearman's rank order correlation analyses unraveled the specific timing of contact acquisition, pointing towards **day 3** as a crucial moment of the hESCs transition between the two states of pluripotency.

4.2.3 Poised enhancers in different contact classes interact with different genes

Given the different trends of PE connectivity observed, I then asked if the PEs of the three different interaction classes contacted different types of genes. Indeed, GO term analysis showed that while genes engaging in interactions with PEs of the UP class are mainly associated with terms such as pattern specification, regionalization, and, more in general, with terms related to development (3e-04 $\leq$ p-value $\geq$ 1e-04), genes contacted within both the DOWN and CONSTANT classes showed a higher enrichment for gene ontology terms mainly associated with more common metabolic processes (e.g. regulation of pH, DNA cell-regulation, cell-cell-adhesion, maintenance of cell polarity, protein complex dissasembly), as shown in **Figure 4.7**.

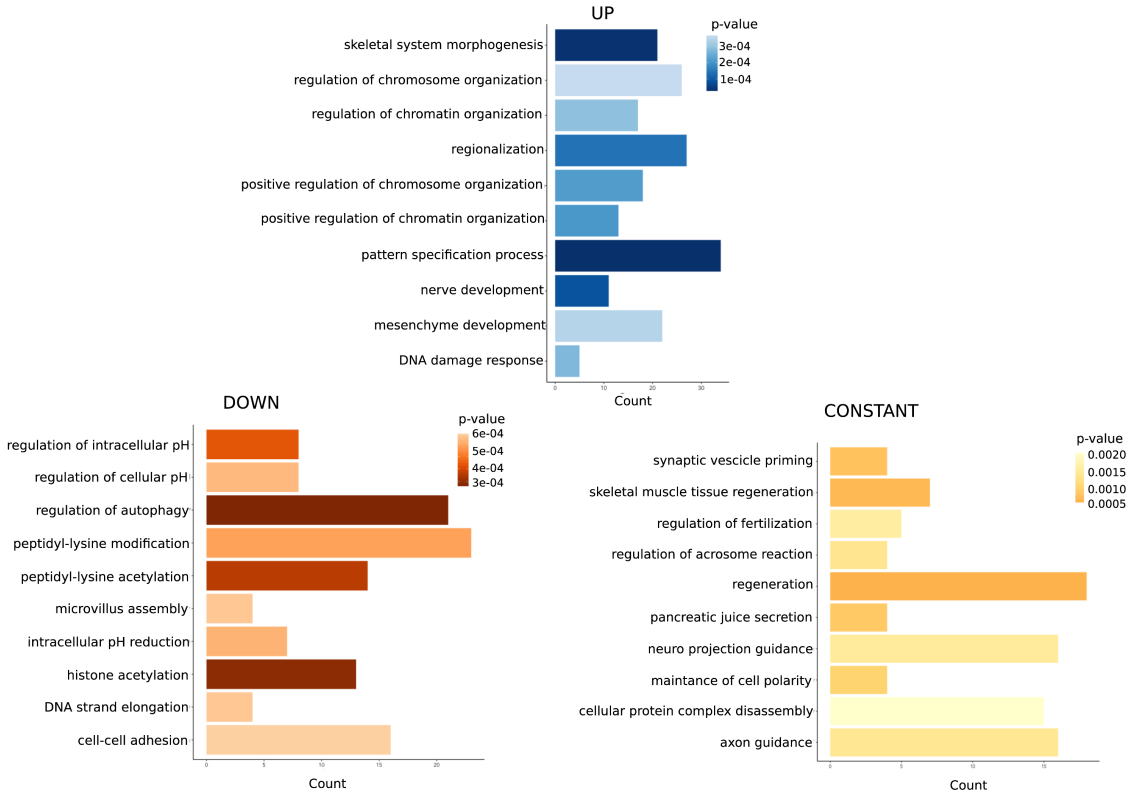


Figure 4.7: **Gene Ontology term analysis for PE interacting genes of the three different classes.** Barplots showing the ten most significant Gene Onotology biological processes terms of genes contacted in each interaction class: UP (blue, 3e-04$\leq$ p-value $\leq$ 1e-04), DOWN (red, 6e-04$\leq$ p-value $\leq$ 3e-04) and CONSTANT (yellow, 2e-03 $\leq$ p-value $\leq$ 5e-04).

The results suggested that, according to the direction of their connectivity associated with each interaction class, PEs can contact different types of genes at different stages of the naïve-to-primed transition, possibly suggesting a more complex regulatory dynamics

of PEs than the one previously proposed.

4.2.4 Interplay between poised enhancer interaction classes and H3K27me3 and H3K4me1 temporal dynamics

In collaboration with Rostovskaya M. (Babraham Institute), I used the recently developed Cut&Tag assay to profile both H3K4me1 and H3K27me3 chromatin binding patterns upon the naïve-to-primed transition [425]. As a control, I first confirmed the enrichment of captured PEs for both H3K4me1 and H3K27me3 in primed hESCs. Indeed, as shown in **Figure 4.8, A and B**, as expected captured regions showed enrichment for both PTMs in the primed hESCs, while the same regions in naïve hESCs only showed enrichment for H3K4me1, but not H3K27me3. Additionally, in line with what has recently been reported, naïve hESCs also displayed a broader distribution of the H3K27me3 mark, which then re-arranged into a more focused configuration and sharper signals in primed hESCs, confirming the major H3K27me3 reorganization previously described between the two pluripotency states (**Figure 4.8, C**).

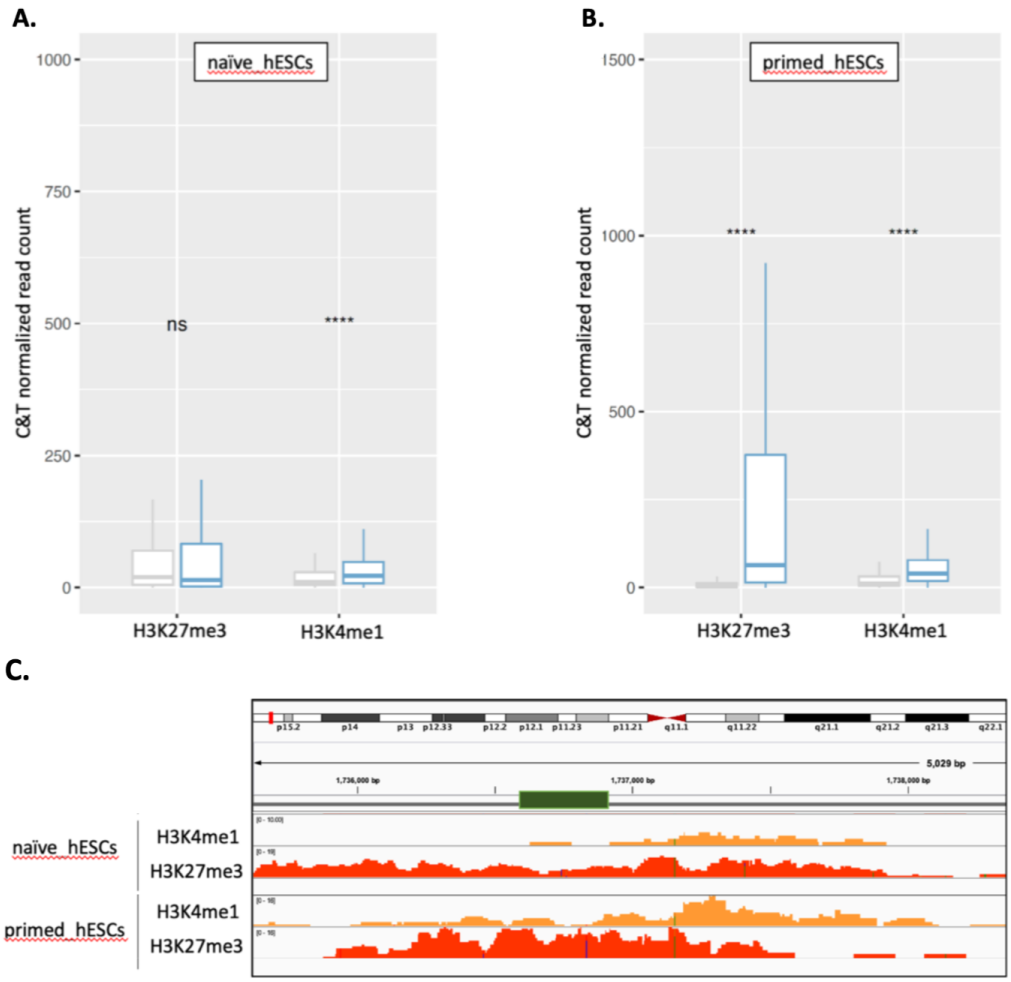


Figure 4.8: **H3K4me1 and H3K27me3 levels at PEs in naïve and primed hESCs. A.** Box plots showing H3K27me3 and H3K4me1 levels (shown as DESeq2 normalized C&T read counts, y-axis) at captured *DpnII* fragments (baits) overlapping candidate PEs in primed (H9) hESCs (blue) compared to randomly selected, "non-captured" *DpnII* restriction fragments (gray). Wilcoxon test, p-value ≤ 0.0001 (****). **B.** Box plots showing H3K27me3 and H3K4me1 levels (shown as DESeq2 normalized C&T read counts, y-axis) at captured *DpnII* fragments (baits) overlapping candidate PEs in naïve (hNES1) hESCs (blue) compared to randomly selected, "non-captured" *DpnII* restriction fragments (gray). Wilcoxon test, p-value ≤ 0.0001 (****). **C.** IGV browser [462] viewpoint showing the distribution of histone PTMs H3K27me3 (red track) and H3K4me1 (orange track) at a typical PE region (dark green bar) included in the PECHi-C capture system in primed (H9) hESCs (top 2 tracks) and naïve (hNES1) hESCs (bottom two tracks).

As can be seen in **Figure 4.9**, H3K4me1 and particularly H3K27me3 levels showed a significant difference between the interaction classes.

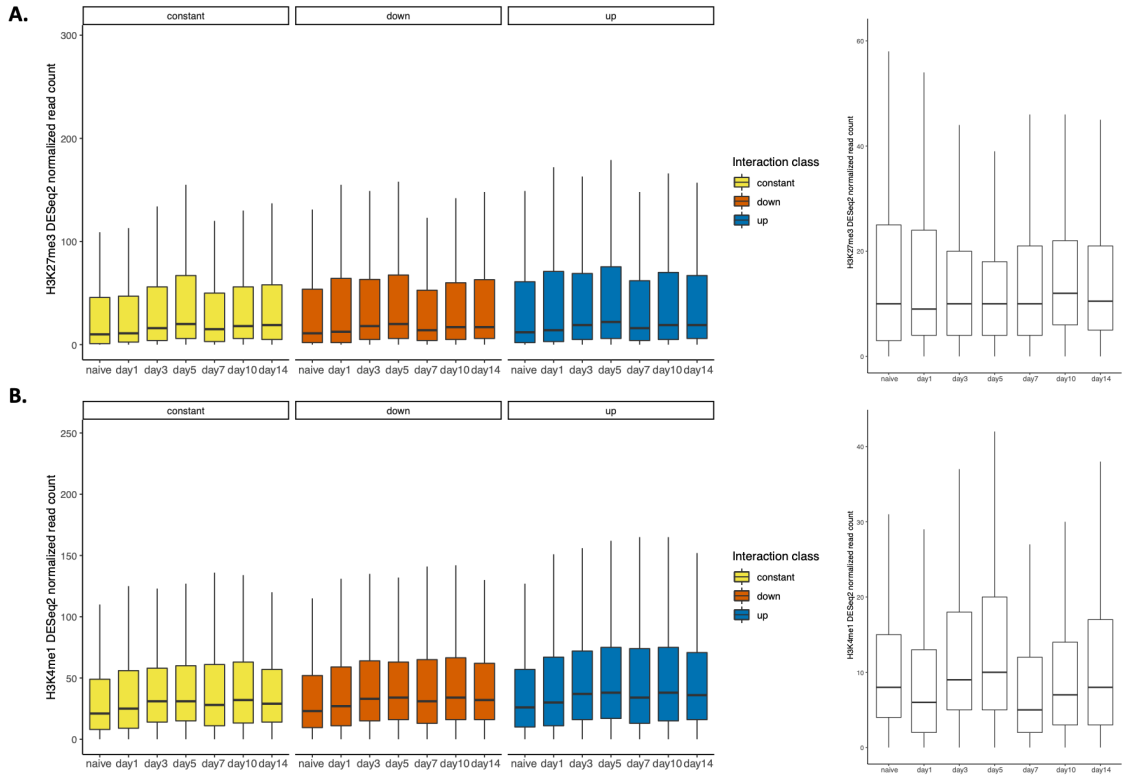


Figure 4.9: **H3K4me1 and H3K27me3 levels at PEs upon the naïve-to-primed transition. A.** Box plots showing H3K27me3 levels (shown as DESeq2 normalized C&T read counts, y-axis) at captured PE regions within the three interaction classes (UP:Kruskal-Wallis p-value ≤2e-16; DOWN:Kruskal-Wallis p-value ≤2e-16; CONSTANT:Kruskal-Wallis p-value ≤2e-16) in hESCs (hNES1) during the naïve-to-primed (pimed H9 hESCs) transition. Box plot on the right (white) shows levels of H3K27me3 for randomly selected, non-baited *DpnII* fragments. **B.** Box plots showing H3K4me1 levels (shown as DESeq2 normalized C&T read counts, y-axis) at captured PE regions of the three interaction classes in hESCs (hNES1) (UP:Kruskal-Wallis p-value ≤2e-16; DOWN:Kruskal-Wallis p-value ≤2e-16; CONSTANT:Kruskal-Wallis p-value ≤2e-16) during the naïve-to-primed (primed H9 hESCs) transition. Box plot on the right (white) shows levels of H3K27me3 for randomly selected, non-baited *DpnII* fragments.

Specifically, while the data showed a gradual increase of levels for both PTMs at PEs compared to background (**Figure 4.9**) in all cases, as cells progressed into the transition, PEs of the UP class displayed higher H3K4me1 and H3K27me3 levels throughout the transition when compared with PEs in the DOWN and the CONSTANT classes (as further confirmed by linear regression analysis, **Figure 4.10, A** and **B**).

Overall, these initial observations suggested that, generally, the bivalent state of enhancers is gradually acquired upon the time course of the naïve-to-primed transition, although PEs within the three interaction classes seemed to display different levels of both

PTMs. Therefore, I set out to formally probe the relationship between interaction classes and H3K27me3/H3K4me1 levels.

In order to describe the contribution of a specific interaction class and time to the acquisition dynamics of the two histone marks, I applied linear regression to model H3K4me1 and H3K27me3 levels (i.e. response variable) defining interaction classes and time as predictor variables and including the sum of such predictor variables in the initial model. Furthermore, to allow the application of the simple linear regression, C&T read counts were transformed using a variance stabilizing transformation.

This model showed the clear association between time and acquisition of both histone marks in all interaction classes (**Figure 4.10, A**, Wald test p-value = $\leq$ 2e-16 for H3K27me3 and p-value = 5.67e-10 for H3K4me1). Next, to incorporate the potential modifying effects of each class on specific timepoints, I further refined the model to include the interaction terms between predictors and, in addition, to include the time as a discrete variable, in order to account for the effect of the interaction classes on H3K4me1 and H3K27me3 levels at specific timepoints rather than across the whole time course (see Methods **section 2.2.8**). The extended model showed that the degree of the increase for both H3K27me3 and H3K4me1 differed between interaction classes. In particular, while for H3K4me1 the interaction term was not significant, for H3K27me3 the UP class significantly affected the magnitude of the increase, "boosting" the levels of H3K27me3 (β = 0.33, Wald test p-value = 1.26 x 10^{-5}). Interestingly, the UP class and day 3 also significantly amplified the increase in H3K27me3 levels at PEs in a synergistic manner(β = 0.25, Wald test p-value = 3.47 x 10^{-3}. **Figure 4.10, B**).

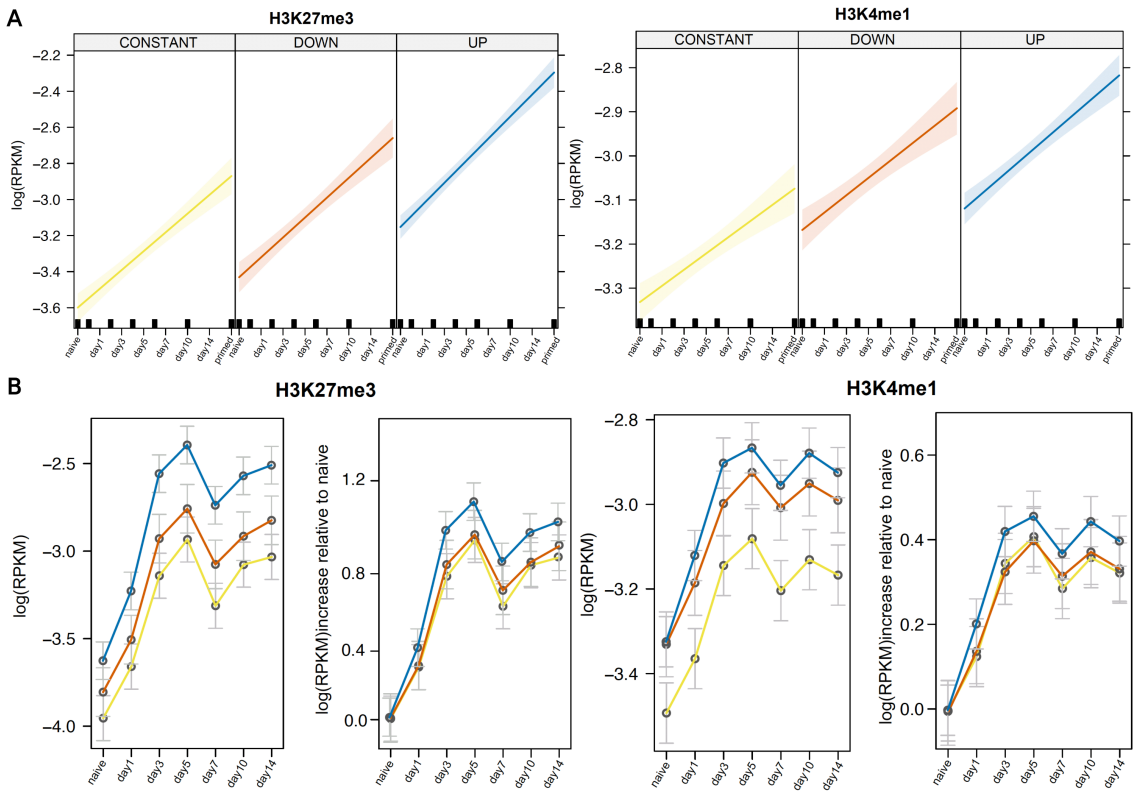


Figure 4.10: **Linear regression model of H3K27me3 and H3K4me1 levels at PEs of the three interaction classes. A.** Effect plots of the linear regression to model the association between H3K27me3, left panel, and H3K4me1, right panel, expressed as logarithm of RPKM (a value of 0.001 was added to prevent infinite values resulting from counts = 0), and time (x-axis) within all interaction classes (CONSTANT = yellow, DOWN = red, UP = blue). **B.** Effect plots of the linear regression model that includes the effect of the interaction terms between predictors (i.e. specific timepoints and interaction classes: CONSTANT = yellow, DOWN = red, UP = blue) for both H3K27me3 levels, left panel, and H3K4me1 levels, right panel, expressed as logarithm of RPKM (a value of 0.001 was added to prevent infinite values resulting from from counts = 0). For both PTMS the increase of H3K27me3 and H3K4me1 levels at each timepoint of the transition relative to the naïve state is shown within each interaction class.

These findings showed that as cells progressed into the transition from the naïve to the primed state of pluripotency, levels of both H3K27me3 and H3K4me1 increased at PEs. While all interaction classes showed a gradual increase of the two PTMs with time, linear regression analysis suggested that PEs of the UP class are characterized by higher starting levels and greater increase of both PTMs, but especially of H3K27me3 (as also shown in **Figure 4.11, B** and **D** box plots). Moreover, it identified **day 3** as a crucial timepoint of the transition for the emergence of both the bivalent state of enhancers (also shown by box plots of **Figure 4.11, A** and **C**) and, in some cases, of their connectivity (i.e. the UP class).

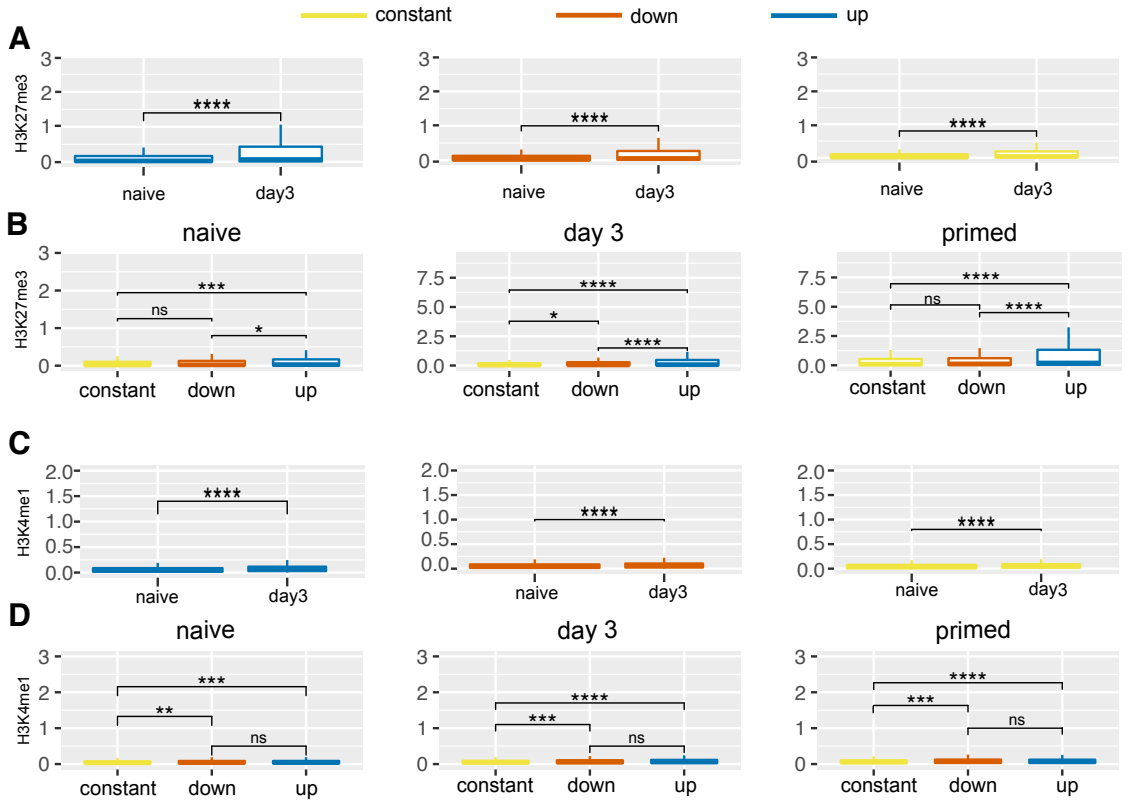


Figure 4.11: **H3K27me3 and H3K4me1 levels at PEs within the three interaction classes. A.** and **C.** Box plots showing H3K27me3 (**A**) and H3K4me1 (**C**) levels in naïve (hNES1) hESCs compared to hESCs (hNES1) at day 3 of the transition at PEs within the three interaction classes. Wilcoxon test p-value ≤0.0001 (****). **B.** and **D.** Box plots comparing levels of H3K27me3 (**B**) and H3K4me1 (**D**) between PEs of the different interaction classes in naïve (hNES1) hESCs, hESCs (hNES1) at day 3 of the transition and in primed (H9) hESCs. Wilcoxon test p-value ≤0.0001 (****), p-value ≤0.001 (***), p-value ≤0.01 (**) or p-value ≥0.05 (ns).

Indeed, although both the DOWN and CONSTANT classes also showed an increase in H3K27me3 levels, but to a lesser extent in comparison to the UP class, such increase was not accompanied by an acquisition of contacts (as seen in the case of the UP class), suggesting that the bivalent state might be necessary, but not sufficient, in the establishment of PE-mediated regulatory networks.

4.2.5 Poised enhancers in different contact classes display different features

The results presented above suggested that the establishment of the bivalent state does not uniformly affect the dynamics of PE chromosomal contacts. Interestingly, recent studies have found that a great majority of annotated PEs are enriched for non-promoter, "orphan" CGIs (oCGIs). It was proposed that oCpGs might play a role in determining PE

responsiveness, activity and contacts [265]. Therefore, I next asked if PEs associated with different interaction classes displayed different CG content.

Making use of all annotated CGIs from UCSC Genome Browser [448], I first looked at the overlap between the PEs included in the PECHi-C capture system and CGIs. Just under one third (32%) of the captured PEs overlapped with CGIs within a 500bp window. These PEs will be referred hereinafter as "CpG positive", while those that are not within a 500bp window from any annotated CGI as "CpG negative". I then asked if the CpG positive PEs showed a preferential association with one of the three interaction classes. Indeed, our results showed that, while PEs of all classes showed a higher association with CGIs than expected (**Figure 4.12, A**), however the association between PEs and CGIs was stronger for PEs within the UP class when compared to the association between CGIs and PEs within the DOWN or the CONSTANT class (**Figure 4.12, B**).

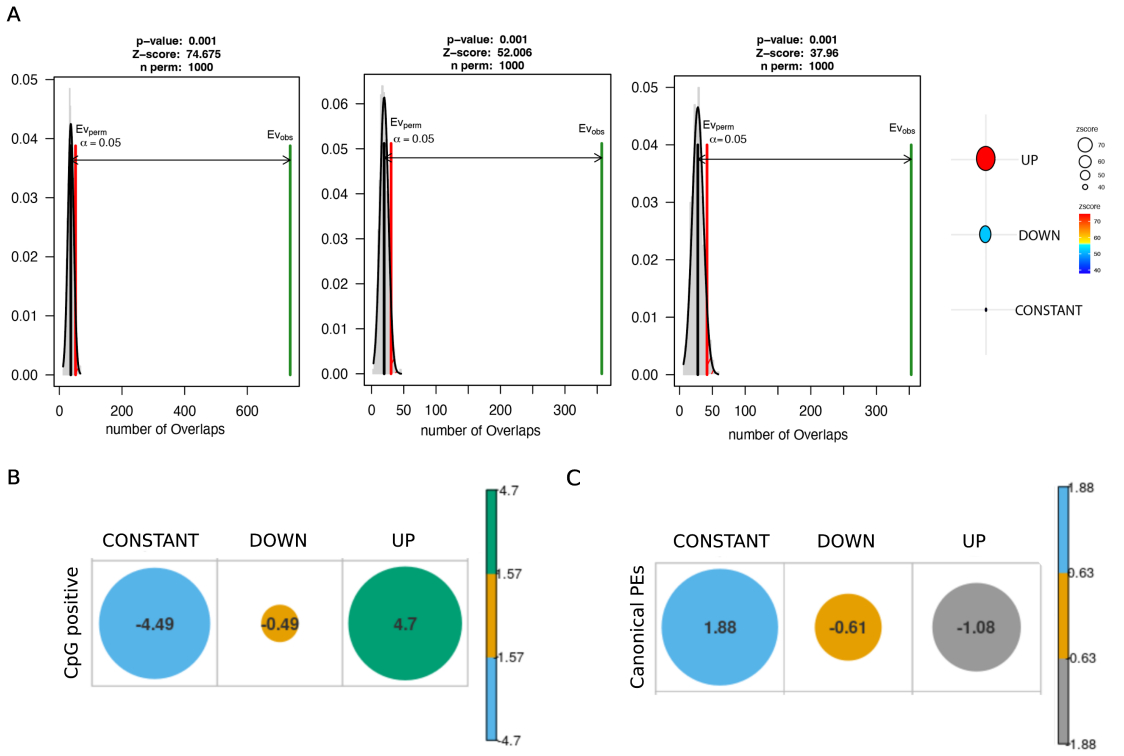


Figure 4.12: **PEs within the UP class are more significantly associated with CGIs but not with canonical PEs. A.** A visual representation showing the results of permutation to assess the association between PEs and CGIs. In grey the number of overlaps of the randomized regions with CGIs, clustering around the black bar (mean. Number of permutations = 1000). In green the number of overlaps of PEs within the different interaction classes (from left to right: UP, DOWN and CONSTANT), which is much larger than expected. The red line denotes the significance limit. Balloon plot on the right hand side shows z-score of the association between CGIs and PEs of each class: 74.675 (UP class); 52.006 (DOWN class); 37.96 (CONSTANT class). See Methods, **section 2.2.11** for more details. **B.** Pearson's standardized residuals plot showing a positive association of CpG-positive PEs with the UP interaction class (p-value = 4.83 x 10^{-3} has been corrected for over-dispersion. See Methods, **section 2.2.11** for more details). **C.** Pearson's standardized residuals plot showing the association between canonical PEs an the three interaction classes (p-value = 3.55 x 10^{-1} has been corrected for over-dispersion. See Methods, **section 2.2.11** for more details).

I then tested whether more generally, the UP class PEs were enriched for canonical PEs (i.e. previously annotated PEs by Pachano T., et al., 2021, [265]). Notably, I found that canonical PEs were not preferentially enriched in either class, suggesting that the features used to define the canonical PE signature are not predictive of the temporal dynamics of their chromosomal contacts upon the naïve-to-primed transition (**Figure 4.12, C**).

Overall, this analysis revealed that PEs of the UP class are also preferentially positioned in proximity of oCpGs, but do not specifically represent a subgroup of the previously annotated canonical PEs [391]. This suggested that additional features, beyond their bivalent signature, their association with CpG-rich regions or their DNA accessibility profile (**Supplementary Figure A.5, Appendix A**), might determine the different trends of PE connectivity observed in my analysis.

4.2.6 Candidate DNA-binding factors determining enhancer poising and connectivity

Given that enhancers can represent landing sites for cell-type and DNA-sequence specific TFs, the idea that PEs within each of the three different interaction classes could differ at a DNA level, perhaps recruiting different TFs with different affinities, seemed plausible.

Using TFs HoCoMoco collection data (see Methods) [443] and making use of the TRAP method for predicting TF affinity for a DNA region based on a biophysical model [444], I computed TFs affinity scores for the underlying DNA sequences of PEs. I then compared affinity scores of any given TF across PEs within different interaction classes and, Kruskal-Wallis one-way analysis of variance identified a total of 113 TFs with significantly different affinities for PEs of the three interaction classes (**Appendix D**). Additionally, I applied the GADEM algorithm for *de novo* motif discovery analysis [446] in order to identify which of the TFs motifs, amongst the 113 hits identified through TRAP analysis, showed a specific enrichment at PEs.

Initial *de novo* motif discovery identified five different predominant motifs to be recurrent at PEs and, interestingly, the top two most enriched motifs showed high CG content (**Figure 4.13**), in line with previous reports [265]. Subsequent Pearson's correlation analysis showed that 63 TFs of the 113 TRAP hits generated an r $\geq$0.3 with at least one of the top 5 motifs identified through GADEM, suggesting that PEs could represent potential binding targets of these TFs, with a different binding profile between PEs within each of the interaction classes. However, the HoCoMoco collection used for this analysis mostly contains information of TF binding sites derived by inference based on ChIP-Seq experiments [443] and, in addition, the differences in DNA accessibility of the regions of interest were not accounted for at this stage.

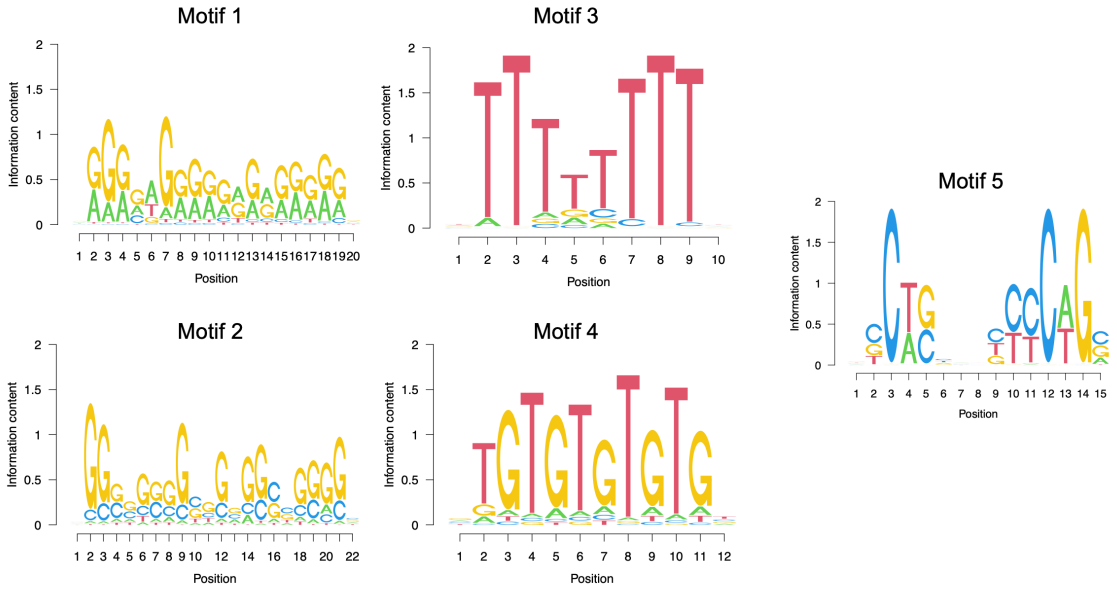


Figure 4.13: **GADEM *de novo* discovery motif analysis identified five predominant motifs at PEs.** GADEM *de novo* motif discovery analysis identified five motifs particularly enriched at PE regions (default p-value cutoff applied of 2e-04).

The PECHi-C and C&T analyses described above highlighted day 3 of the naïve-to-primed transition as a crucial timepoint, when a sharp increase in both PEs connectivity and PTMs levels occurs. Interestingly, in line with these findings, previously published RNA-seq analysis studies identified a specific sub-group of genes that showed a peak of expression taking place at day 3 of the transition [424] (**Figure 4.14, A** and **B**).

I used the RNA-seq data (GSE123055) from this study to ask whether this group included any of the 63 TF candidates identified in the motif analysis, which would suggest their potential involvement in the establishment of PE-mediated contacts. I first re-analyzed the data and confirmed the presence of the five specific gene expression patterns previously described, including the sub-set of genes with a peak of expression at day 3 (**Figure 4.14, A** and **B**).

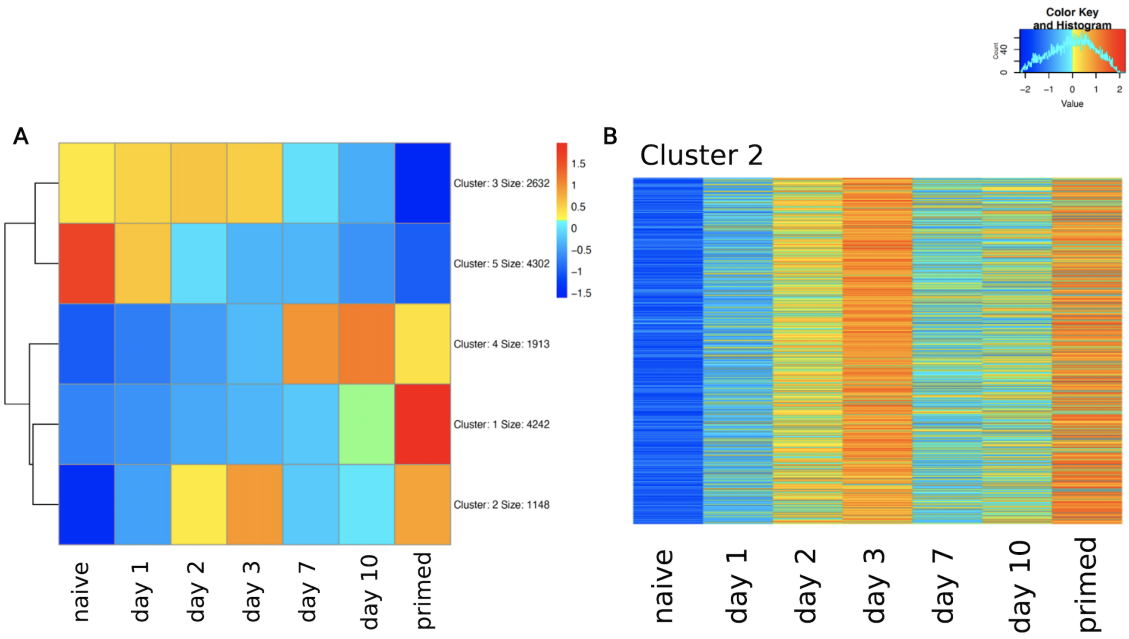


Figure 4.14: **Differential gene expression analysis upon the naïve-to-primed transition in hESCs. A.** Differential gene expression analysis of publicly available RNA-seq data confirmed the presence of five specific clusters of genes showing differential expression patterns as cells transition from the naïve (hNES1) to the primed (H9) state (timepoints analyzed are: naïve (hNES1), day 1 (hNES1), day 2 (hNES1), day 3 (hNES1), day 7 (hNES1), day 10 (hNES1), primed (H9) hESCs) [424]. Heatmap plotted using *z-score* of normalized read counts. **B.** Zoomed-in heatmap of cluster 2, representing the set of genes showing a peak of up regulation specifically at day 3 of the transition. Heatmap plotted using *z-score* of normalized read counts. RNA-seq analyzed are publicly available at GSE123055 [424]

Two TFs mapped to the cluster of genes showing an up-regulation on day 3: **PBX2** and **ZBTB14**. Both PBX2 and ZBTB14 consensus motifs (shown in **Figure 4.15, A** and **B**) showed a similarity score of $r = 0.6$ and $r = 0.3$ (Pearson's correlation coefficient), respectively, with *de novo* motifs identified at PE regions, which suggested the possibility for these two factors to bind at PE regions. Notably, for ZBTB14, but not for PBX2, the presence of the DNA-binding motif at PEs has also been confirmed by an alternative *de novo* motif discovery analysis using the HOMER software [447] (**Figure 4.15, C**). Additionally, both TFs showed different affinities for PEs of the three interaction classes, with ZBTB14 showing a higher affinity for PEs within the UP class compared with the PEs of the DOWN and, particularly, the CONSTANT classes. On the other hand, PBX2 showed a significantly higher affinity for PEs in the DOWN and the CONSTANT classes when compared with the PEs in the UP class (**Figure 4.15, A** and **B**).

These analyses suggested that ZBTB14 and PBX2 may bind at PEs and, owing to their temporal patterns of expression, could be implicated in the establishment of PEs regulatory network.

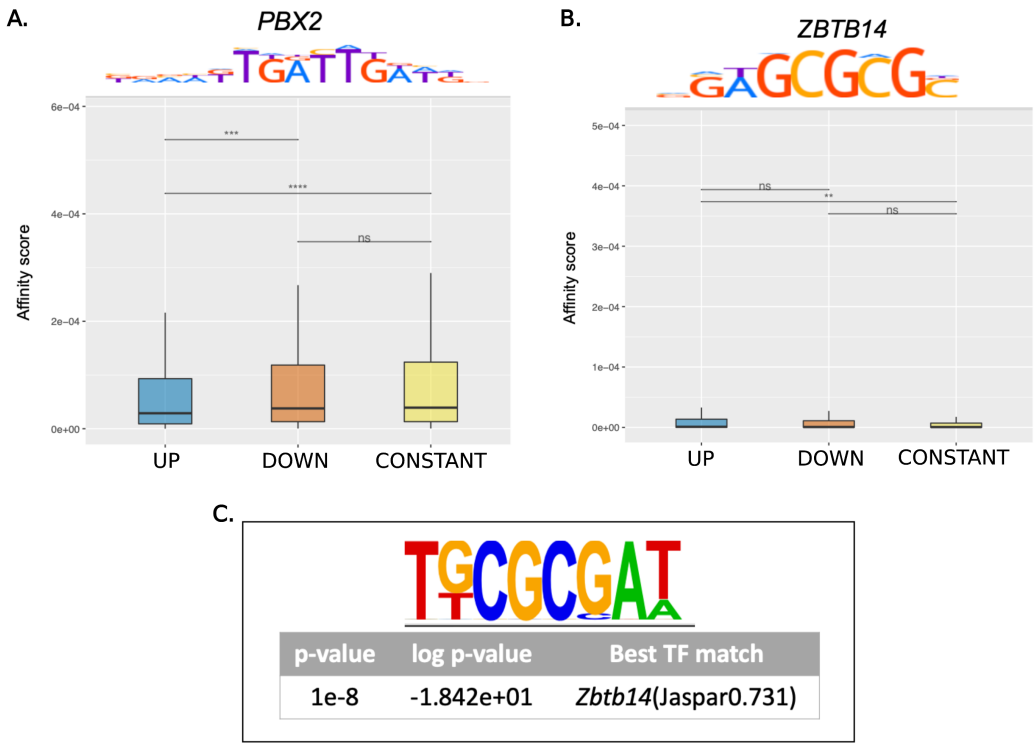


Figure 4.15: **PBX2 and ZBTB14 affinity scores for PEs within the three different interaction classes. A.** and **B.** Box plots comparing the affinity scores (y-axis, min-max normalized affinity scores) of PBX2 and ZBTB14 for PEs within the three interaction classes. Dunn's test p-values $\leq$ 0.0001 (****), p-value $\leq$ 0.001 (***), p-value $\leq$ 0.01 (**), p-value $\geq$ 0.05 (ns). The consensus motif for both TFs is shown. **C.** *De novo* motif identified at PEs of the UP class through HOMER motif discovery analysis with a high similarity to ZBTB14 consesus motif (p-value=10^{-9}).

In addition to the sequence-specific factors identified in the analysis above, core chromatin co-factors may play a role in mediating enhancer poising and 3D contacts. The main example is given by PcGs and the numerous evidence showing that PcG-bound domains establish long-range interactions in ESCs and, in particular, that PEs establish contacts with their target genes in a PRC2-dependent manner [264]. In addition, recent studies have identified DPPA2 and DPPA4 as factors involved in the establishment of bivalency at promoters in hESCs [414].

DPPA2 and DPPA4 are heterodimerizing nuclear proteins (Developmental Pluripotency Associated 2 and 4, respectively) involved in the regulation of zygotic genome acti-

vation (ZGA)-associated transcription. Interestingly, DPPA2/4 also show binding of non-ZGA genes, in particular at bivalent promoters [463, 413, 464, 465, 466]. Given their described role in the establishment and maintenance of bivalency at promoters, I then asked if DPPA2/4 could be implicated in the regulation of the bivalent state at PEs as well, and perhaps play a role in mediating PE-promoter interactions in the context of pluripotency.

Using DPPA2/4 unpublished ChIP-seq data from Rugg-Gunn lab (Babraham Institute, Cambridge), I looked at the binding profile of the two factors at PEs, with a specific focus on PEs within the main three interaction classes.

First, I confirmed the enrichment of both DPPA2 and DPPA at the PEs included in our PECHi-C assay. Indeed, both factors were significantly enriched at PEs in both naïve and primed hESCs (**Figure 4.16, A**), although displaying higher levels in cells in the primed state (**Figure 4.16, B**, top panel). Interestingly, the analyses revealed that PEs in naïve cells are pre-bound by both DPPA2 and DPPA4 although being characterized by lower levels of H3K27me3, as shown in **Figure 4.16, B** (bottom panel). This suggested that the binding of both factors might precede the acquisition of higher levels of H3K27me3 at PEs, hence the recruitment of PcGs, inferring their potential role in the establishment of bivalency and PE-mediated regulatory network.

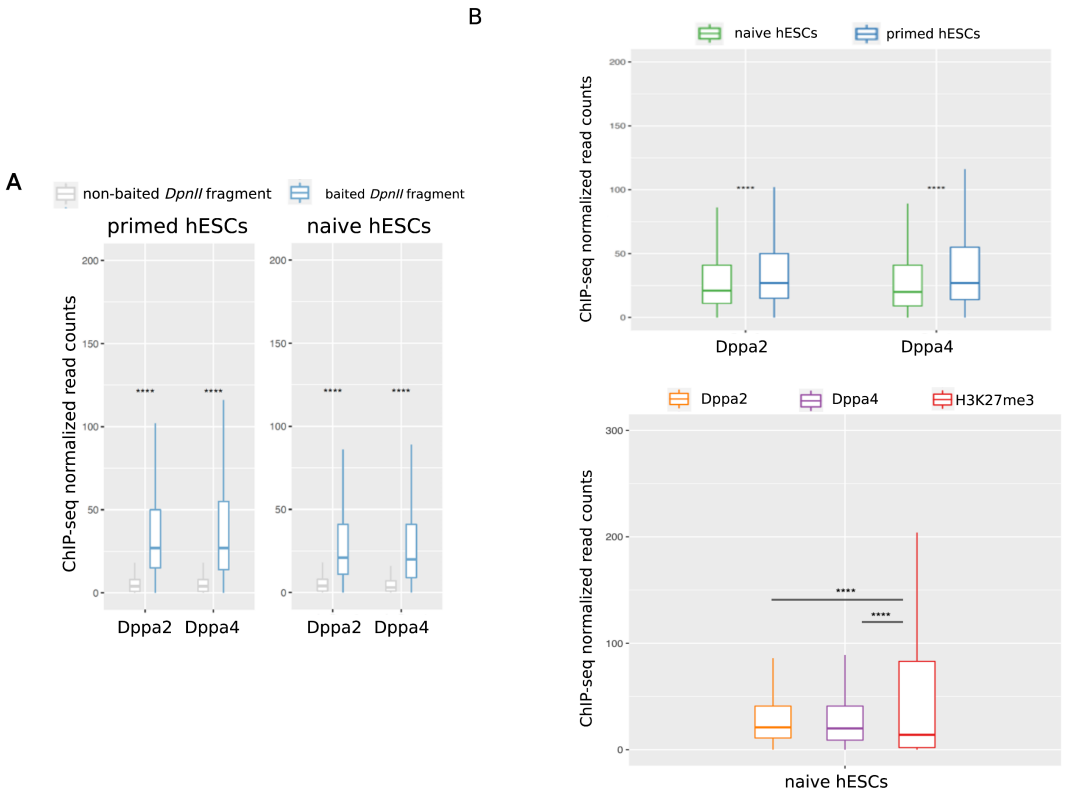


Figure 4.16: **DPPA2 and DPPA4 binding at PEs. A.** Box plots showing the comparison of Dppa2/4 levels (y-axis, ChIP-seq normalized read counts) between baited *DpnII* fragments (blue) and non-baited *DpnII* restriction fragments (gray) in naïve and primed hESCs. Wilcoxon test p-value ≤ 0.0001 (****). **B.** Box plots comparing levels of Dppa2/4 at baited *DpnII* fragments (y-axis, ChIP-seq normalized read counts) between naïve (hNES1) (green) and primed (H9) hESCs (blue), top panel. Wilcoxon test p-value ≤ 0.0001 (****). Bottom panel box plots comparing the levels of Dppa2/4 (orange and purple, respectively) with H3K27me3 levels (red) at baited *DpnII* fragments in naïve (H9) hESCs. Wilcoxon p-value ≤ 0.0001 (****).

Given the presence of the different trends of PEs connectivity, I next asked if PEs within a specific interaction class were specifically enriched for either DPPA2 or DPPA4. Indeed, both factors were preferentially associated with PEs of the UP class, in naïve and primed hESCs alike, while in both the DOWN and CONSTANT classes only a small percentage overlapped with DPPA2/4 bound regions, with a higher percentage of PEs showing lower levels for both factors (**Figure 4.17**).

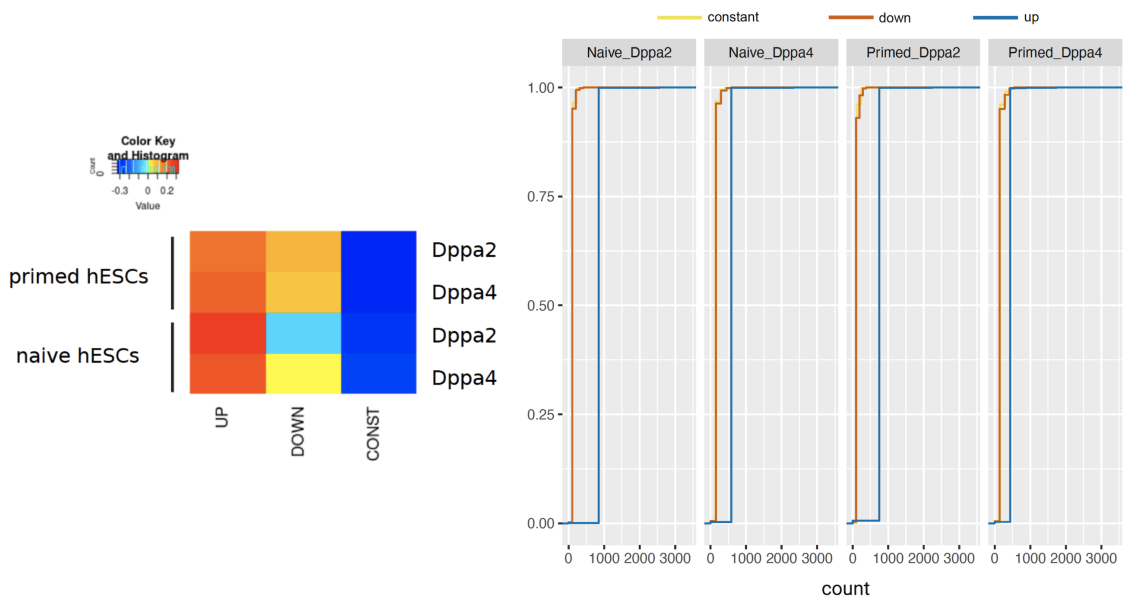


Figure 4.17: **DPPA2/4 enrichment at PEs across the three interaction classes.** Left panel showing the heatmap of log odds ratio of Dppa2/4 peak enrichment at PE regions within the three interaction classes. Empirical cumulative distribution function (right panel) showing the proportion of PEs (y-axis) with a specific Dppa2/4 read count (x-axis) within the three interaction classes (UP = blue; DOWN = red; CONSTANT = yellow).

Overall, the analysis presented in this section showed that PEs within each interaction class can have different affinities for specific TFs, leading to the recruitment of different TFs and co-factors. Specifically, it revealed DPPA2, DPPA4, PBX2 and ZBTB14 as candidate factors potentially involved in the establishment and maintenance of bivalency at PEs, alongside PcGs. In particular, the observation of a specific enrichment at PEs within the UP interaction class for DPPA2, DPPA4 and ZBTB14 suggested their potential role in the establishment of PE-mediated interactions, providing an additional explanation to the three different trends of PE connectivity that arise upon the naïve-to-primed transition.

Further experimental validation of the analyses presented in this section will be key to establish or exclude a potential role for these factors in the emergence of the bivalent signature of PEs and their contacts as cells progress from the naïve to the primed state of pluripotency.

4.2.7 Setting up the tools for testing the effects of enhancer activation with an inducible CRISPRa system (iCRISPRa)

While initial evidence supported the role of PEs as crucial regulatory elements for the expression of genes upon differentiation [264], the importance and the functional meaning of their defining bivalent nature still remains unclear. To uncover the functional role of PE bivalency, I sought to establish a CRISPR/Cas9-mediated perturbation system that will represent a valuable tool for the systematic perturbation of the the chromatin state of candidate PEs.

I set out to test an iCRISPRa system in human induced Pluripotent Stem Cells (iPSCs) for the artificial activation of genes by targeting promoters and enhancers (henceforth referred to as iCRISPRa iPSCs). Originally established in the Kampmann lab (UCSF), the iPSCs line stably expresses dCas9-VPH under the regulation of DHFR (dihydrofolate reductase) which causes regular proteosomal degradation of the dCas9-VPH. The addition of trimethoprim (TMP) counteracts DHFR-mediated degradation and stabilizes the expression of the dCas9-VPH, which can be targeted to promoters or enhancers of interest through specific sgRNAs (a schematic of the construct if shown in **Figure 4.18, A**). In the context of PEs, the advantage for the use of an inducible dCas9-mediated chromatin perturbation system is given by the possibility to target PEs of interest at any significant point as cells transition from the naïve to the primed state of pluripotency. For example, the perturbation of candidate PEs before the acquisition of their bivalent state or prior to the emergence of their contacts could represent a powerful strategy to gain further insights on the functional role of the acquisition of the bivalent signature and how it mediates the emergence of PE-mediated regulatory circuitry.

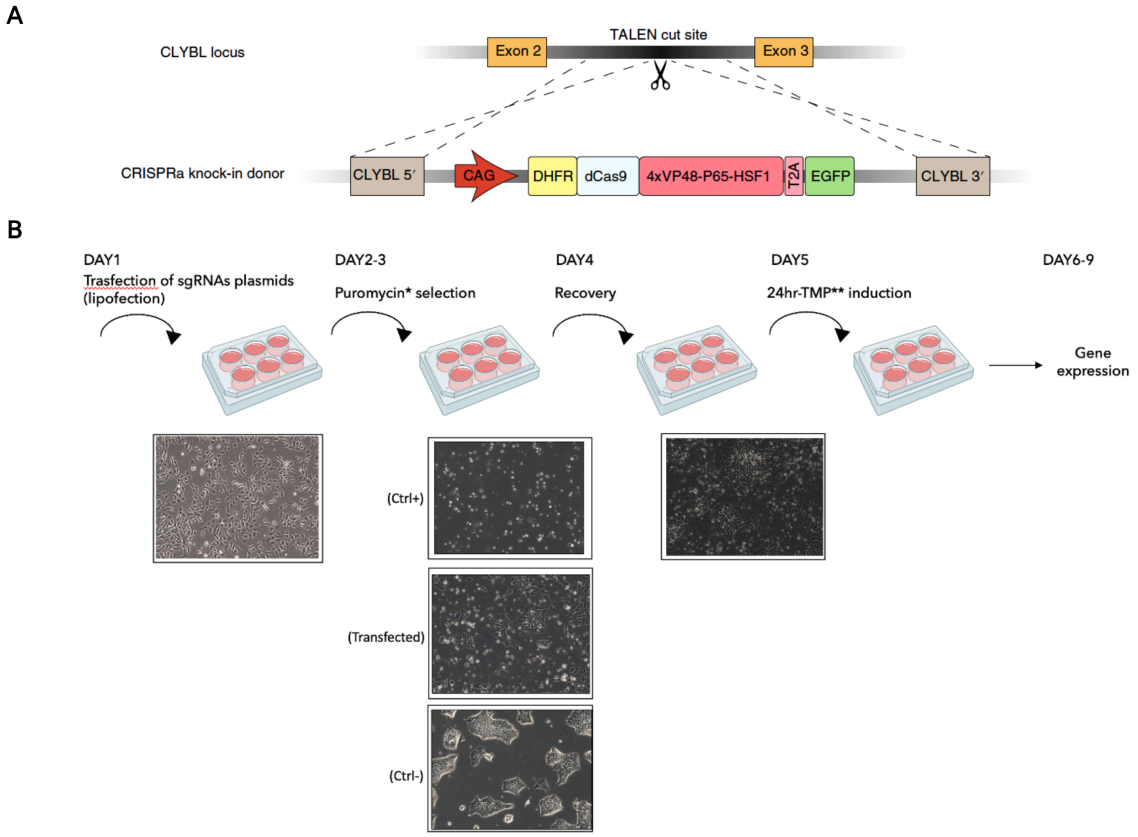


Figure 4.18: **Schematic of the inducible CRISPRa system in iPSCs. A.** Construct for the generation of the inducible CRISPRa cell line (iPSCs, WTC11). DHFR-dCas9-VPH (DHFR, dehyfrofolate reductase; dCas9, catalytically dead Cas9; VPH, activator domains containing 4× repeats of VP48, P65 and HSF1), were stably integrated into the CLYBL safe harbor locus using a TALEN-mediated knock-in. Figure from Tian, R., et al, 2021 [423]. **B.** The experimental design for the activation of target promoters and enhancers through the iCRISPRa system in iPSCs. Ctrl+: iPSCs treated with puromycin; Ctrl-: iPSCs without puromycin; Transfected: iPSCs transfected with sgRNA-bearing plasmids making sucessfully transfected cells resistant to puromycin. *puromycing = 2μg; **TMP = 20μM.

To test the iCRISPRa system, iPSCs were transfected with sgRNAs targeting known promoters and enhancers. Specifically, I targeted the promoters of *NEUROD1*, *CXCR4* and *GATA1* genes using previously tested sgRNAs as positive controls [423, 426]. Moreover, to test the system not only for promoters, but for *cis*-regulatory regions as well, an annotated *GATA1* enhancer, which did not display an enrichment for any of the main PTMs (i.e. H3K4me3, H3K4me1, H3K27ac, H3K27me3) in hESCs, was also targeted using previously validated sgRNAs [426].

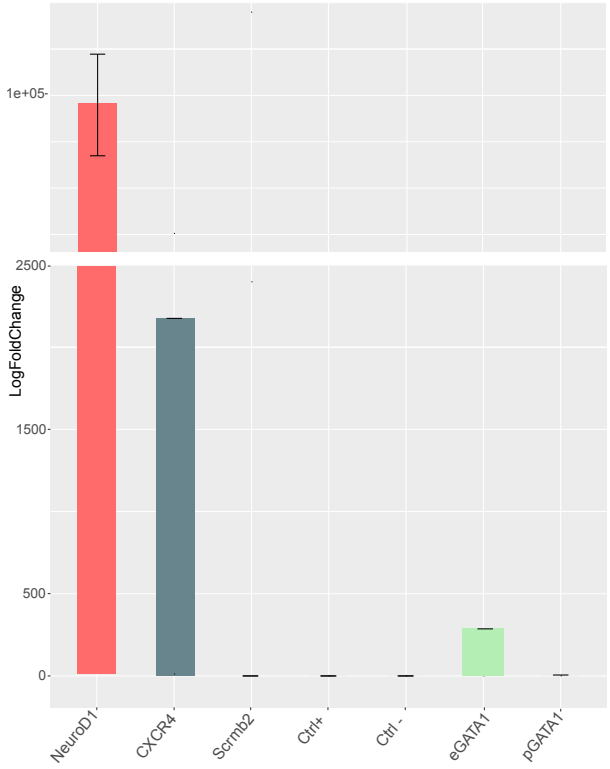


Figure 4.19: **Inducible CRISPRa (iCRISPRa) for activation of *cis*-regulatory elements in iPSCs.** Validation of iCRISPRa activity via qPCR in iPSCs (WTC11). qPCR quantification of the relative fold change of *NEUROD1* (red), *CXCR4* (dark cyan) and *GATA1* (light green) mRNA levels in iCRISPRa iPSCs (WTC11) expressing sgRNAs targeting the promoter of *NEUROD1*, *CXCR4*, *GATA1* (i.e. pGATA1) and a known enhancer for *GATA1* (i.e. eGATA1) compared to a non-targeting control sgRNA (Scrmb2) in the presence or absence of TMP (Ctrl+ and Ctrl-, respectively), which stabilizes the DHFR degron (error bars represent standard deviation. n = 3 biological replicates for *NEUROD1* and *CXCR4* and n = 2 biological replicates for *GATA1*.

qPCR assays showed successful up-regulation of *NEUROD1* and *CXCR4* by targeting their promoters (up to 140,000 and 2,000 fold, respectively). *GATA1* up-regulation was also successfully achieved by targeting its enhancer (up to 200 fold). However, targeting *GATA1* promoter did not result in the activation of its expression (**Figure 4.19**). The lack of activation observed when targeting *GATA1* promoter could lie in the choice of the sgRNAs (it is plausible that different sgRNAs have different capacity to achieve induction or repression of gene expression) and the number of sgRNAs used to target each region in this experimental setting. Indeed, due to technical limitations, while ten previously reported strong sgRNAs were used to target *GATA1* enhancer [426], only two sgRNAs were used for the *GATA1* promoter, perhaps decreasing the chances of achieving the induction of the expression of the target gene.

Nevertheless, the successful up-regulation of *NEUROD1*, *CXCR4* and *GATA1* showed that activation of gene transcription could confidently be achieved by targeting both known promoters and enhancers with suitable sgRNAs in iPSCs through the iCRISPRa system described above.

Recent CRISPR-based screening studies have reported the combination of CRISPR-dCas9 mediated perturbation with RNA fluorescence in situ hybridazation (FISH) coupled with sorting of fluorescent-labelled cells (FACS), henceforth referred to as RNA-Flow FISH, a sensitive single-cell method for the detection of downstream transcriptional effects after CRISPR-mediated perturbation [467]. Therefore, I sought to optimize the RNA Flow-FISH methodology in hESCs and its coupling with our established iCRISPRa system in iPSCs. In the context of PEs, this methodology will allow to detect potential subtle effects after perturbation of PEs and, in addition, will be more amenable to identify suitable sgRNAs in a higher-throughput screening setting.

I first optimized the RNA Flow-FISH protocol for its use in hESCs by comparing the expression profiles of *OCT4* in primed hESCs and neuroepithelium differentiated hESCs. As shown in **Figure 4.20, A**, RNA Flow-FISH successfully detected RNA expression in hESCs and, furthermore, a shift in the *OCT4* expression profile between primed hESCs and differentiated cells could be observed as early as day 5 into differentiation.

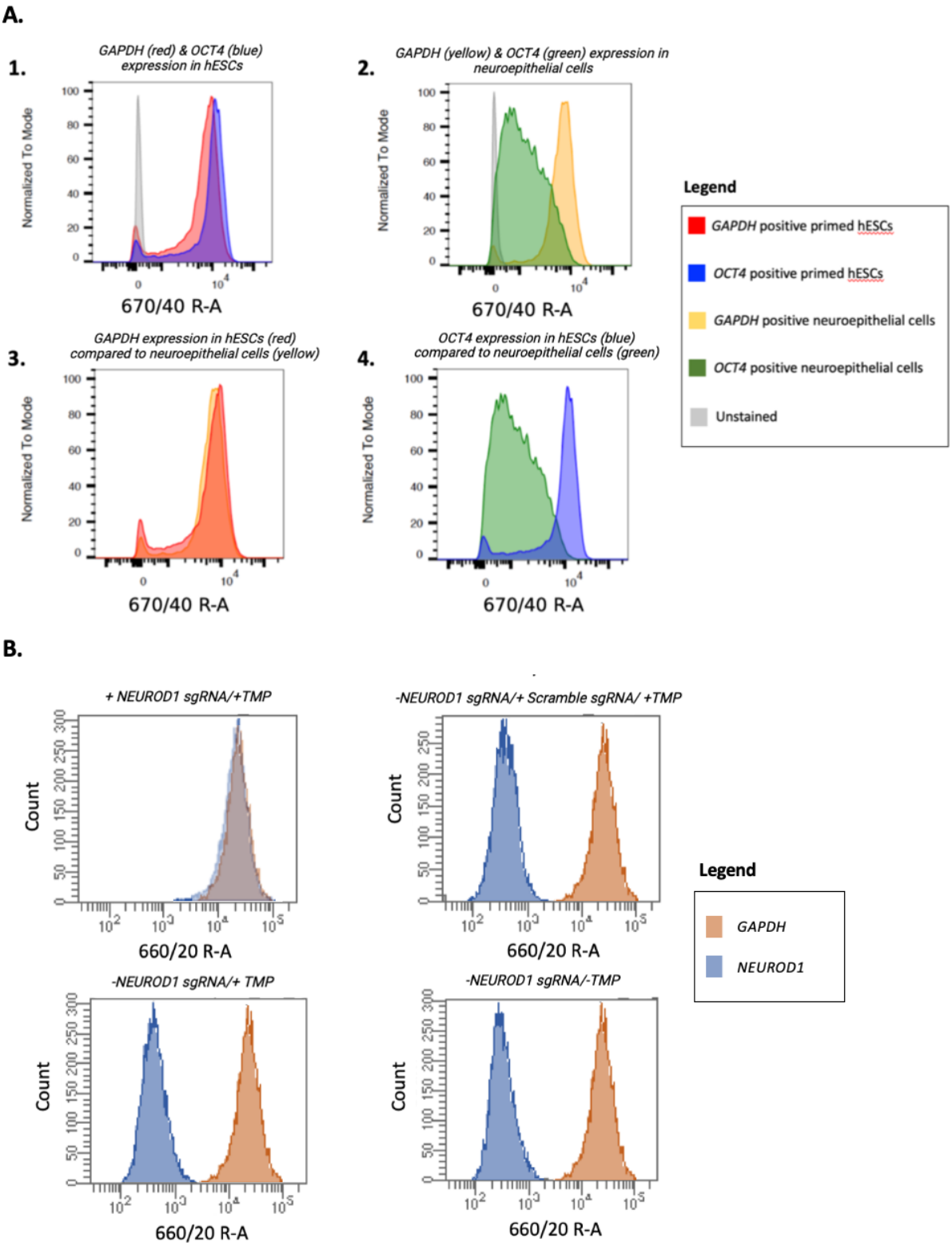

A.
1.
GAPDH (red) & OCT4 (blue) expression in hESCs
Normalized To Mode
670/40 R-A
2.
GAPDH (yellow) & OCT4 (green) expression in neuroepithelial cells
Normalized To Mode
670/40 R-A
Legend
GAPDH positive primed hESCs
OCT4 positive primed hESCs
GAPDH positive neuroepithelial cells
OCT4 positive neuroepithelial cells
Unstained
3.
GAPDH expression in hESCs (red) compared to neuroepithelial cells (yellow)
Normalized To Mode
670/40 R-A
4.
OCT4 expression in hESCs (blue) compared to neuroepithelial cells (green)
Normalized To Mode
670/40 R-A
B.
+ NEUROD1 sgRNA/+TMP
Count
660/20 R-A
-NEUROD1 sgRNA/+ Scramble sgRNA/ +TMP
Count
660/20 R-A
Legend
GAPDH
NEUROD1
-NEUROD1 sgRNA/+ TMP
Count
660/20 R-A
-NEUROD1 sgRNA/-TMP
Count
660/20 R-A

Figure 4.20: **RNA Flow-FISH allows to detect subtle gene expression changes in hESCs and can be coupled with CRISPRa induced changes in gene expression.** Validation of RNA FISH-Flow in hESCs (H9) and iPSCs (WTC11) for detection of mRNA levels. **A. A. Detection of *GAPDH* and *OCT4* expression in hESCs (H9) and neuroepithelial differentiated cells by RNA Flow-FISH. Panel 1**: *GAPDH* (red) and *OCT4* (blue) expression levels detected by flow cytometry in H9 hESCs (unstained population of cells shown in gray). **Panel 2**: *GAPDH* (yellow) and *OCT4* (green) expression levels detected by flow cytometry in H9 hESCs neuroepithelial differentiated cells (unstained population of cells shown in gray). **Panel 3**: comparison of *GAPDH* expression levels, detected by flow cytometry, between H9 hESCs (red) and H9 hESCs 5 days into neuroepithelial differentiation (yellow). **Panel 4**: comparison of *OCT4* expression levels, detected by flow cytometry, between H9 hESCs (blue) and H9 hESCs 5 days into neuroepithelial (green). Counts are normalized to mode to account for the different number of events recorded (y-axis). Intensity of fluorescence detected expressed on a logarithmic scale (x-axis. 670/40 R-A = Alexa647). **B. Detection of the up-regulation of *NEUROD1* expression in iCRISPRa iPSCs (WTC11).** Clockwise: *NEUROD1*(light blue) and *GAPDH* (orange) expression levels detected by flow cytometry in iCRISPRa iPSCs (WTC11) expressing sgRNA targeting *NEUROD1* promoter (+*NEUROD1* sgRNA) upon addition of TMP (+TMP); *NEUROD1* (light blue) and *GAPDH* (orange) expression levels detected by flow cytometry in iCRISPRa iPSCs (WTC11) expressing Scramble sgRNA, targeting a random non-regulatory genomic region, (+Scramble sgRNA) upon addition of TMP (+TMP); *NEUROD1* (light blue) and *GAPDH* (orange) expression levels detected by flow cytometry in iCRISPRa iPSCs (WTC11) not expressing the sgRNA targeting *NEUROD1* promoter (-*NEURO* sgRNA) in absence or presence of TMP ($\pm$TMP). Intensity of fluorescence detected expressed on a logarithmic scale (x-axis. 670/40 R-A = Alexa647). The same number of events were acquired for each sample (counts on y-axis). Analysis of flow-cytometry data we performed in FlowJo, by the Babraham Institute FlowCytometry facility.

I then tested the coupling of the methodology with the iCRISPRa system. Successful specific up-regulation of *NEUROD1* gene expression was observed upon TMP addition in iCRISPRa iPCSs (WTC11) expressing a sgRNA targeting the *NEUROD1* promoter (*NEUROD1* sgRNA) (**Figure 4.20, B**), compared to the lack of up-regulation of *NEUROD1* expression in cells that were not expressing the *NEUROD1* sgRNA or were expressing, instead, a scramble sgRNA (targeting a random, non -regulatory region of the genome). These results confirmed the successful coupling of RNA Flow-FISH with the iCRISPRa system in iPSCs for specific induction of the expression of genes of interest (in all samples the expression levels of *NEUROD1* were compared to housekeeping *GAPDH* expression levels, **Figure 4.20, B**).

The work presented in this section described the successful set up of a iCRISPRa system for the perturbation of promoters and enhancers in iPSCs, which can be coupled with a single-cell readout for gene expression complementary to qPCR assays. The coupling of iCRISPRa and RNA Flow-FISH has the potential to be a powerful tool to perturb candidate PEs and gain insight on their functional role in the regulation of expression programs upon lineage commitment.

4.3 DISCUSSION

This chapter presented the profiling of the emergence of the poised state of enhancers and their connectivity upon naïve-to-primed transition in hESCs.

Devising the PECHi-C approach has enabled me to track PEs contacts genome-wide upon the naïve-to-primed transition revealing three predominant patterns of PEs connectivity: UP, DOWN and CONSTANT. Interestingly, PEs that showed an increased connectivity with time, namely the UP class, showed a correlation between the poising of the enhancer regions and the emergence of their contacts, preferentially interacting with developmental genes. However, this did not appear to be the case for the two remaining classes, the DOWN and the CONSTANT class, suggesting the presence of different "subgroups" of PEs that can potentially be part of different regulatory mechanisms.

A key question remains about PEs' functional role during early embryogenesis and, specifically, about the relevance of their distinctive bivalent nature. Here, I presented preliminary work aimed to set up an inducible CRISPR-Cas9 based approach for the artificial perturbation of the chromatin state of candidate PEs in hPSCs. This approach represents a valuable tool to perturb PEs and their bivalent signature and uncover functional aspects of the role of PEs in pluripotency and cell-fate decision.

4.3.1 A potential role for poised enhancers in the naïve-to-primed transition in hESCs

When PEs were first described in ESCs, it was hypothesized that they were part of an elegant mechanism to prime genes for their prompt expression upon differentiation, as in the case of other BDs such as bivalent promoters. Indeed, evidence based on 4C and CRISPR-based disruption studies of PEs supported such hypothesis, showing that PEs represented major regulatory regions and contacted their target genes in ESCs in a Polycomb-dependent

manner, before becoming fully active upon differentiation [233, 264].

Notably, several PCHi-C studies [468, 278] showed that Polycomb-mediated long-range interactions undergo a significant re-organization when hESCs transition from the naïve to the primed state of pluripotency, offering the first indication that, perhaps, PE-mediated interactions could undergo a similar degree of re-wiring between the two states. Having access to an efficient system to transition naïve hESCs into the primed state [424] allowed me to profile the timing of occurrence of both the poising and the connectivity of enhancers, as a first step to elucidate the role of PEs in pluripotency. Moreover, the PECHi-C approach presented in this chapter, gave the advantage to profile genome-wide contacts that specifically involve PEs, as opposed to capturing annotated promoters alone, providing a first insight into PE-specific regulatory networks.

Through the PECHi-C approach I was able to identify three predominant patterns of PE contacts upon the naïve-to-primed transition in hESCs: the **UP**, **DOWN** and **CONSTANT** interaction classes. Given the observed abundance of bivalency in primed hESCs and based on the evidence of a greater PcG-dependent interaction network in the primed state of pluripotency [278], the UP class described an anticipated pattern, showing a gain of PE connectivity upon transition. Less obvious was the presence of the DOWN and CONSTANT classes of interactions, whereby PE-mediated contacts are either lost or retained as cells proceed into the transition, suggesting the presence of a more complex regulatory mechanism. The observation of different connectivity trends could suggest a role for PEs, not only at later stages of differentiation, but also in the naïve-to-primed transition in hESCs. However, the data presented in this chapter are not sufficient to fully answer this question and systematic dissection of the function of PE regions, for example through perturbation assays, will be crucial to establish a possible mechanism in support of this hypothesis.

A fundamental feature of PE-mediated interaction is the well-known role of Polycomb group proteins (PcGs). PcGs represent one of the most studied chromatin mediated mechanisms for gene repression, especially in the context of development and maintenance of cell identity. In addition to their activity as chromatin modifying enzymes, it is now clear that PcGs play an important role in mediating chromatin looping interactions between regions marked by H3K27me3 that can often reside in different TADs and can be established in a cohesin-independent manner [128]. Both PRC1 and PRC2 have

been implicated in the establishment of interactions between PcG-bound loci, in particular between PcG-bound promoters and PEs. However, it is becoming clearer that PcG complexes might not act alone in mediating chromatin looping interactions between PcG-bound regions and it is plausible to hypothesize that a specific combination of DNA-specific binding factors might be involved in directing PcGs to specific sites in the genome [469, 470]. Based on this possibility, it could be hypothesized that the different trends of PE connectivity revealed by the PECHi-C data, might not be exclusively mediated in a PcG-dependent manner, but could be established by different mechanisms whereby PcG complexes can act in combination with other factors or, in some cases, might not be implicated at all.

There is evidence that suggests that PRC1 and PRC2 can bind at different "types" of bivalent domains. For example, it has been shown that while PRC1 tends to bind at bivalent promoters that display higher conversation across species and are generally transcriptionally repressed, PRC2 tends to be found mostly at CpG-rich bivalent regions. Interestingly, the preferential binding of PRC2 to specific CpG-rich regions also seems to correlate with specific motif content of such regions, along with their H3K27me3 and transcriptional levels [469, 471]. The data presented in this chapter revealed that different subgroups of PEs could be, indeed, distinguished based on features like: CpG-content, underlying DNA-specific motifs and H3K27me3 and H3K4me1 occupancy ratio. It can be speculated that, despite their general shared bivalent signatures, these additional features might determine the recruitment of different PcG complexes and/or different DNA-binding factors. Specifically, I found that PEs of the UP class significantly associated with CpGs compared to PEs in the DOWN and the CONSTANT classes. Their preferential association with CpG-rich regions could be linked to the acquisition of higher levels of H3K27me3 with subsequent recruitment of PcGs and establishment of interactions with their target genes. On the other hand, the absence of CpGs, or rather the lower association with PEs with the DOWN and the CONSTANT classes, can be associated with the acquisition of lower levels of H3K27me3 observed at these regions.

Interestingly, despite their significant association with CpG-rich regions, PEs of the UP class do not seem to specifically represent previously described canonical PEs, which have also been found in proximity of CpG-rich regions [265]. It is worth noting that the definition of canonical PEs is not based solely on the co-presence of H3K4me1 and H3K27me3, which is instead the definition used to devise the PECHi-C approach here described, but these regions are also defined based on the presence of p300 and on DNA accessibility (i.e.

ATAC-seq) [391]. The observation that canonical PEs do not show a specific connectivity trend upon transition, together with the finding of a clear association between UP class PEs and CpGs, seems to suggest that the definition of PEs used in the PECHi-C approach results in a less stringent classification of PEs, potentially more inclusive of different subgroups of PEs.

It has been suggested that the association of PEs with CpG-rich regions could be responsible of the increased responsiveness and connectivity of PEs, leading to the speculation that CpGs might act as tethering elements [265]. This hypothesis would be in agreement with the correlation that I observed between PEs in the UP class, found in proximity of CpG-rich regions, and the temporal dynamics of their connectivity. The notion that enhancer connectivity and enhancer activity might be driven by distinct regions has been recently gaining ground, with recent high resolution 3D chromatin conformation studies supporting this hypothesis [472]. Due to the use of the *DpnII* enzyme (four-cutter), the resolution of PECHi-C data presented in this chapter is greater than Capture Hi-C datasets generated with the more commonly used *HindIII* enzyme (six-cutter). However, the degree of resolution given by *DpnII* it is not sufficiently high to clearly disentangle this scenario for the specific case of PEs and their associated CpG-rich regions.

An additional aspect of CpG-associated PEs is the observation that they tend to preferentially establish contacts with promoters that are also CpG-rich [265]. Additionally, in most cases, enhancers and promoters are known to share chromatin features such has PTMs profiles. Based on these observations, it would be interesting to probe such concordance between PEs of the UP class and their target promoters to understand further the differences underlying the PECHi-C interaction classes. Indeed, a possible concordance of chromatin features between PEs and their target genes could support the idea that PE-dependent interactions within the different classes are mediated by different factors, alone or in combination with PcGs. This might ultimately lead to a scenario whereby different types of PE contacts may counteract each other establishing a more complex gene regulatory mechanism.

Interestingly, the analyses presented in this chapter highlighted **day 3** of the transition as a crucial tipping point, whereby the sharpest increase in the emergence of both contacts and histone marks (i.e. H3K4me1 and H3K27me3) could be observed. It is known that during the naïve-to-primed transition, cells at day 3 show the most significant changes at

expression levels, as well. Moreover, the relevance of day 3 seems to be recapitulated *in vitro*: prior to day 3, cells can easily revert back to their naïve state by simply switching culturing conditions. Once past day 3, cells seemingly enter a more "stable" state and are set on their course to become primed. Indeed, work from Rostovskaya, M. et al, 2019 showed that during the first 3 days of the transition, cells down-regulate a subset of naïve factors, such as *KLF4* and *TFCP2L1*, which coincides with the reduction in the ability of the cells to reform naïve colonies [424]. The finding that day 3 appears to be important also for the emergence of PE bivalency and, in some cases, their connectivity (i.e. UP class), further suggests a role for PE-mediated regulatory network in pluripotency. It would be plausible to hypothesize that the different trends of interactions observed, and a possible re-wiring of interactions between classes (for example a PEs that switches between target promoters, as cells transitions between the two pluripotency states, namely going from the DOWN class to UP class), could be necessary to determine the correct exit from the naïve state, "locking" the cells onto their progression into the primed state of pluripotency.

Overall, the work in this chapter presented the tracking of the emergence of the bivalent state of enhancers and their connectivity over the time course upon the naïve-to-primed transition and evidence of how their poised state may affect their connectivity. Making use of a unique window on early embryogenesis, our results suggest potential new mechanisms of action of PEs, beyond simply priming of genes for prompt expression upon differentiation, suggesting a role for PEs during the transition between the two states of pluripotency.

While, undoubtedly, chromatin state profiles and connectivity are important aspects to shed light on the regulatory networks of PEs, alone they are not sufficient to paint a complete picture. The results here described, alongside recent evidence, suggested the possible involvement of additional factors that could regulate the emergence of PEs and their contacts network. Further integration of this data and additional genome-wide chromatin profiling data will be crucial to uncover further details and, perhaps, confirm novel roles of PEs in early embryogenesis.

4.3.2 Other players in the establishment of bivalency

Could the different PE connectivity patterns be explained by dissimilarities at the DNA level, where the presence of different DNA-binding motifs can determine the recruitment

of different factors and co-factors at PEs? Indeed, through motif discovery analyses I identified PBX2 and ZBTB14 as potential candidate factors. For both TFs, a potential involvement in cell differentiation and developmental processes has been described. Particularly, *PBX2* gene is part of the TALE/PBX homeobox family, a class of factors and co-factors classified as architects of the body plan during development [473], while *ZBTB14* encodes for a TFs that can act as both activator and repressor and it has been implicated in the regulation of the formation of neural tissues in vertebrates. Based on the enrichment of their motifs at PEs and their characteristic affinity patterns for the different interaction classes, it could be speculated that the recruitment of different TFs and co-factors can lead to different PE-mediated regulatory mechanisms, hence leading to the observations of the three different trends of PE connectivity, despite their overall shared bivalent signature.

Of particular interest is ZBTB14, a TF that binds to 5'-d(GCC)(n)-3' trinucleotide repeats at promoters and which preference for CpG-rich regions it is in agreement with its higher affinity with PEs of the UP class. Moreover, in a recent CRISPR-KO screen study, ZBTB14 has been implicated in the regulation of germline genes. Interestingly, this study showed that the absence of *ZBTB14* drives transcriptional responses that lead to a reactivation of a 2-cell like state signature [474], placing ZBTB14 on the scenes as a potential factor involved in the regulation of pluripotency and prompting for further exploration of its role in the context of PEs.

Recently, two additional factors have been implicated in the establishment and maintenance of bivalency at promoters: DPPA2 and DPPA4 [413, 414]. Based on the idea that, generally, bivalent promoters are more likely to interact with bivalent enhancers (i.e. PEs), given the concordance between their chromatin state, these findings suggested that DPPA2/4 might play a role in the context of PEs as well. ChIP-seq analysis presented in this chapter showed that PEs in the UP class are significantly enriched for these two factors in naïve and primed cells alike. Moreover, these enhancer regions seem to be already bound by both DPPA2 and DPPA4 in the naïve state, which led to hypothesize that they could act as signposts to mark specific regions that will acquire higher levels of H3K27me3 and H3K4me1 as cells transition into the primed state of pluripotency.

Both factors have been implicated in the regulation of proper gene expression at later stages of differentiation, as knock out ESCs fail to efficiently differentiate due to loss of bivalency at developmental promoters [413]. This is in agreement with our observations that PEs in the UP class, which showed an enrichment for both DPPA2 and DPPA4, contact

genes preferentially involved in differentiation and development. Taken together, these initial observations seem to suggest a role for DPPA2 and DPPA4 in the establishment of bivalency both at bivalent promoters and PEs alike. Interestingly, different studies also described the possibility that both DPPA2 and DPPA4 could interact with members of the PcGs complexes. Indeed, recent findings suggested that rather than acting solely as transcriptional repressors, PcGs might cover a more versatile role according to their interaction "partners" [475], further supporting the hypothesis that DPPA2/4, in some cases, might cooperate with PcGs in the establishment of PE-mediated regulatory network in ESCs [476, 414].

Overall, this chapter presented data that, in alignment with recent evidence, suggested that Polycomb might not act alone in the establishment of bivalency and PE-mediated connectivity. Multiple other factors, acting alone or in combination, might contribute to define PE functional and temporal dynamics.

The results presented here suggest a scenario where the PE regulatory network could be important for the transition from the naïve state to the primed state of pluripotency. However, at this stage the analyses are mostly based on correlation approaches which do not provide an exhaustive explanation for the presence of the three interaction classes, nor provide clarity on which might be their functional and biological meaning.

In addition, it is important to consider some technical limitations of our approach. For example, although candidate PEs included in the PECHi-C capture system have been identified at the best of our abilities, it cannot be fully excluded that some of the bivalency observed might be the result of population heterogeneity and could, potentially, add noise to the data generated. Furthermore, our PECHi-C data analysis only looks at pair-wise interactions, which at this stage prevented us from looking at potential hubs of PEs, whereby the three classes are not separate events, but might show a greater degree of inter-connectivity than we were not able to appreciate in our analysis. Going forward, it would be valuable to integrate the data presented in this chapter in network-based approaches to explore, for example, PE connectivity beyond the pair-wise context, with strategies similar to the one recently described for PCHi-C data by Chovanec, P., et al, 2021 [278].

4.3.3 Towards elucidating the functional role of poised enhancers in pluripotency

Despite studies in recent years have provided evidence that certainly confirmed the regulatory function of these regions, the functional role of the poising of enhancers remains unclear. Being able to manipulate and artificially perturb the chromatin state of these regions, rather than disrupting them, represents a valuable strategy to shed light on the importance of the poising of enhancers. This chapter presented the preliminary work towards the establishment of an inducible CRISPR-activation system in human iPSCs to perturb the chromatin state of potential candidate PEs.

The validation of the iCRISPRa system presented in this chapter, and its coupling with RNA Flow-FISH for single-cell detection of mRNA expression [477], will enable to initially perturb candidate PEs in primed hESCs and gain first insight into the different mechanisms of different sub-groups of PEs. For example, in parallel perturbation of UP class PEs and CONSTANT class PEs will allow to compare the response of the two classes of PEs to artificial activating stimuli and determine whether the two classes of PEs are equally able to convey the positive signal to their target genes, inducing their expression. Moreover, with the evidence gathered on the timing of the emergence of their bivalent chromatin state and their connectivity, this approach could represent a valuable tool to intervene before the acquisition of PEs' bivalent signature and, possibly, before the emergence of their contacts, uncovering functional aspects of the role of PEs in pluripotency and cell- fate decision.

In addition, coupling CRISPR-based assays with the RNA Flow-FISH method would offer the advantage to perform high-throughput CRISPR-based screen [467] to identify sgRNAs suitable for the activation of a specific candidate enhancer. This strategy could also represent a valid alternative readout for final gene activation. Indeed, given its single-cell nature, it would be possible to distinguish between different "degrees" of gene activation and, likewise, detect eventual population heterogeneity. In addition, it would represent a useful tool for the detection of gene expression in those instances of enhancers perturbation within regulatory hubs, for example, whereby perturbing "secondary" enhancer regions might not achieve the significant levels of gene activation usually observed upon promoter perturbation.

Overall, the inducible CRISPR-activation system validated in this chapter will enable chromatin perturbation assays of promoters and enhancers in human iPSCs. Coupled with the RNA Flow-FISH approach, which I validated in hESCs and iPSCs, this system presents a powerful tool for the high-throughput screens of sgRNAs targeting candidate PEs. Moreover, it paves the way for targeting multiple candidate PEs within the same cell, in order to investigate scenarios beyond pair-wise contacts and to address the potential inter-connectivity between PEs hubs, perhaps validating the presence of different subgroups that achieve regulation of gene expression through different mechanisms.

4.3.4 Conclusion

In conclusion, in this chapter I explored the poising and connectivity dynamics of PEs upon the naïve-to-primed transition in hESCs. Although further experimental and computational investigation is needed to shed light on the functional and biological meaning of the three interaction classes, our data suggest that PEs represent, not only a mechanisms to ensure proper gene activation upon differentiation, but might also be implicated in the correct progression of ESCs between the different stages of pluripotency.

Additionally, in line with recent studies [470, 413, 414], my analyses indicate the involvement of other factors and co-factors in the establishment of both bivalency and PE connectivity, beyond the established role of PcGs. In particular we identified PBX2, ZBTB14, DPPA2 and DPPA4 as potential candidates, although further validation will be necessary to confirm or exclude their participation in the regulation of PEs regulatory network and its role in pluripotency.

Finally, this chapter presented the preliminary work for the artificial perturbation with CRISPR-based assays of selected candidate PEs. While it comes with its challenges, the system here described will represent a valuable tool to elucidate the functional role of the poising of enhancers through targeted chromatin perturbations, as an alternative to the more disruptive approaches described so far.

5 General Discussion

The work presented in this thesis focused on enhancers in their poised state, characterized by a bivalent chromatin signature marked by the active H3K4me1 and the Polycomb-associated H3K27me3 repressive mark. As hESCs go through a change in their "identity", switching from a ground-state of pluripotency (the naïve state) to the primed state that is more responsive to differentiation, a global redistribution of PcGs and H3K27me3 occurs. [278]. The findings presented in this thesis showed that PEs and their regulatory circuitry also undergo major reorganization as cells transition between the two states of pluripotency.

In this thesis, PE-mediated contacts were tracked genome-wide as hESCs transitioned from the naïve to the primed state using a PECHi-C capture system devised to specifically enrich for 3D-chromatin interactions which directly involved PEs. Over the last decade, Capture Hi-C (CHi-C)-based methodologies have been key for the exploration of promoter-enhancer crosstalk dynamics in a plethora of different biological contexts and cells types. The low-cell number CHi-C protocol used in this thesis represented a great advantage for the generation of 3D-chromosome conformation data in the context of early embryogenesis. In particular, the use of Tn5-mediated tagmentation for the generation of NGS-ready libraries significantly reduced library processing times, facilitating PECHi-C data generation upon the time course of the naïve-to-primed transition. Indeed, Tn5-mediated tagmentation has been employed in numerous NGS-based methodology such as DNA accessibility and chromatin profiling assays (i.e. ATAC-seq, Cut&Tag) [478]. Nevertheless, the optimization process presented in chapter 3 highlighted how Tn5-mediated tagmentation is a critical step, whose conditions must be precisely controlled in order to robustly generate high quality final libraries. Here, the CHi-C protocol was optimized to obtain the best possible degree of valid pairs and capture efficiency in hESCs, representative of "true" interactive pairs and enrichment for interactions of regions of interest, respectively. In general, while I highlighted crucial steps responsible for

the initial variability observed in Hi-C/CHi-C libraries, the protocol is still characterized by some degree of variability that is currently being addressed. For example, in the specific case of hESCs one important aspect could be represented by their chromatin state. A more open chromatin state could potentially result in a higher degree of over-digestion by the restriction enzyme of choice and the subsequent stochastic insertion of adaptors mediated by Tn5 might highlight these properties in the final library by generating a higher proportion of invalid fragments.

As hESCs transition between the two states of pluripotency, the high number of cells necessary to perform the most commonly used Hi-C and Capture Hi-C protocols, which normally require ≥ 30 million cells, would represent a significant limitation. Therefore, the work towards the further refinement of the previously developed miniaturized Hi-C and Capture Hi-C protocol to process samples as small as 100K cells [452] laid the necessary foundation for the generation of PECHi-C data to track the emergence of poised enhancers in hESCs during the time course of the naïve-to-primed transition.

The method described in this thesis will join the ranks of other refined Hi-C and Capture Hi-C protocols (e.g. OmniC, Micro-C and Capture Micro-C) [20, 289] that enable the profiling of the 3D chromatin structure of cells at increased resolution. Moreover, while the approach explored in this thesis was mainly fine tuned for the analysis of data generated using a four-cutter enzyme such as *DpnII*, it represents a valuable approach for the downstream analysis and the identification of 3D-contacts from data at different degree of resolution.

Since first evidence of PEs in ESCs emerged, great effort has gone into the characterization of these regions and the investigation of their functional role. One of the leading hypothesis is that PEs "prime" genes for rapid activation upon differentiation. Indeed, Cruz-Molina, S., et al 2017 found that, in some cases, the bivalent state of a subset of PEs gets resolved upon differentiation in favor of an active state. They showed that the deletion of PEs significantly compromised the expression of their target genes as mESCs were differentiated in anterior neural progenitor cells (AntNPCs) [264]. However, it cannot be excluded that the bivalent signature can also be resolved in favor of a repressive state to ensure rapid suppression by PcGs of the target genes. Indeed, in support of the idea that PcG-bound loci can act as silencers, Ngan, C.Y., et al 2020 showed that deletion of PRC2-bound loci leads to de-repression of their target genes in mESCs [479]. In addition to their dual role in mediating prompt activation and/or repression of gene expression,

recently bivalent regions have been described as "roadblocks" for DNA methylation in mESCs. McLaughlin, K., et al 2019 showed that, perturbation of DNA methylation in mESCs can affect PcG-mediated chromatin interactions [480], suggesting a close interplay between DNA-methylation and PcG-mediated 3D chromatin conformation. This led to the hypothesis that bivalency at bivalent promoters and at PE regions could represent a mechanism to protect specific regions from terminal silencing. Terminal silencing could ultimately lead to disease development if not properly counteracted, as it has been hypothesized for some specific cases of cancer [481].

While PEs were initially characterized as a main feature of ESCs and mainly involved in ensuring proper differentiation, it is now recognized that they also play a role in other biological contexts and, as previously mentioned, can mediate different regulatory mechanisms. Indeed, through the PECHi-C approach I identified three different trends of PE connectivity, the **UP**, the **DOWN** and the **CONSTANT** classes, that suggested the possibility of the presence of different sub-classes of PEs. In particular, the findings presented in this thesis suggest that, according to their chromatin features (e.g. H3K4me1/H3K27me3, CpG content, DNA-specific TF motifs), subgroups of PEs could recruit distinct DNA-binding factors and co-factors, determining different PE connectivity dynamics as cells transitioned from the naïve to the primed state of pluripotency. For example, the data showed that PEs of the UP class are significantly associated with CpG-rich regions, possibly determining the higher levels of H3K27me3 observed at these regions. Typically, CpG-rich domains are highly associated with PRC2 in ESCs [469], therefore it could be hypothesized that the establishment of PE-mediated interactions of the UP class specifically relies on PRC2.

In general, PcGs are fundamental players in the establishment of 3D looping interactions, contributing to both gene repression and activation [482, 483, 108, 468, 484, 485]. Whether the presence of PcG factors at enhancers ultimately determines the activation or the repression of their cognate genes likely depends on the DNA-specific factors and co-factors they interact with. For example, the presence of PRC2 could inhibit the deposition of H3K27ac and mediate repression, whereas, in other cases, the cooperation between PcGs and activating TFs could promote the binding of CBP/p300 and/or Mediator, leading to the eviction of PcG components and promoting the activation of gene expression. In addition, there remains a possibility that at least some PE-mediated contacts do not depend entirely on PcGs, but can also rely on other chromatin looping factors (e.g. CDK-Mediator) [470]. The data presented in this thesis highlighted the presence of different

sub-classes of PEs characterized by distinct underlying features and different interaction dynamics, which suggest that PEs might combine different mechanisms of action to exert their regulatory role. Our results suggest that, along with their recognized role to ensure correct gene activation upon differentation, PEs might also be necessary at earlier stages of embryogenesis to ensure the establishment of the different states of pluripotency, similarly to what recently suggested by Kumar, B., et al., 2022 and Zijlmans, DW. et al., 2022 [486, 487, 488]. In particular, the importance of **day 3** of the naïve-to-primed transition highlighted by PECHi-C and C&T data, in line with findings by Rostovskaya, M., et al., 2019 [424] that day 3 is a timepoint of critical changes during the transition, further suggests a role for PEs in regulating the correct exit from the naïve state and placing the cells onto the path to the primed state. However, the correlative nature of the findings at this stage does not allow to provide a direct answer to this question. Indeed, while this study mainly looked at histone PTMs (i.e. H3K4me1 and H3K27me3) and 3D-contacts, in the future it would be useful to measure the occupancy of PcGs at PE specific regions directly. Moreover, the targeted depletion of PcGs components would be a valuable strategy to gain further insight into their explicit involvement in the establishment of PE-mediated interactions. Likewise, whether different TFs and DNA-binding factors at PEs, such as the candidates identified in this thesis (i.e. *PBX2*, *ZBTB14* and/or *DPPA2/4*), cooperate with either one of the PcG complexes, mediating different regulatory mechanisms, is a hypothesis that will need to be experimentally validated.

An additional key question concerning PEs is the functional role of their distinctive bivalent state. As previously mentioned, the deletion of candidate regions that bear the PE signature validated their regulatory nature [264]. However, one of the downsides of this disruptive approach is that it does not allow the analysis of the functional role of the "poising", as it inevitably removes its bivalent nature together with the regulatory region itself. Targeted epigenetic perturbation techniques, such as CRISPR-mediated activation or inhibition, could represent a more suitable approach to specifically probe the possible role of the dual chromatin state of PEs [489].

To pave the way for these experiments, here I established an iCRISPRa approach in iPSCs with the aim to perturb the chromatin state of candidate PEs and probe the functional role of their bivalent signature. Based on the timing of the emergence of PE bivalency and connectivity, namely day 3 of the transition, this system would allow to artificially activate enhancers prior to the acquisition of their poised state and their contacts and study the

consequences on gene expression and PE-mediated contacts of such perturbation. Furthermore, while the findings presented in this thesis showed that the acquisition of bivalency and PE-mediated contacts, in some cases, seem to take place in parallel, it does not unequivocally clarify the causal relationship of the two events. Perturbing the chromatin state prior to day 3 has the potential to shed some light on such causality between acquisition of bivalent state and contacts.

However, the iCRISPRa system used in this thesis presented some limitations. The reprogramming from the primed to the naïve state proved inefficient for the iPSCs cell line employed (P. Rugg-Gunn's lab, personal communication), therefore it is not possible to perturb candidate PEs at different timepoints during the transition in this specific setting. A valid alternative strategy to overcome such limitations could be attempting the "de-poising" of PEs by perturbing PcG complexes themselves. Although it is possible to generate naïve and primed hESCs knock-out lines for different components of PcG complexes, the disruption of PcG complexes would not be limited to disrupting PEs only, but would largely affect PcG-mediated gene regulation as a whole, leaving results harder to interpret. Interestingly, a more targeted approach recently developed uses a system whereby a PRC2-specific inhibitor can be targeted at regions of interest and achieve precise reduction of H3K27me3. Such approach would allow to specifically perturb the bivalent signature at PEs, thus removing the background noise that would result from the general disruption of PcGs [490].

Overall, this thesis provided further insights into the PE-mediated regulatory circuitry and its emergence during early embryogenesis. These findings highlighted the presence of different subgroups of PEs, suggesting that their different regulatory mechanisms could be driven by specific features of PE regions, predictive of their different trends of connectivity. PEs have been described as important regulatory regions in ESCs necessary for prompt activation of genes at later stages of differentiation. However, findings in this thesis suggest that PEs might also be important at earlier stages of embryogenesis, promoting the correct transition from the ground state of pluripotency to the primed state through the fine-tuning of gene expression profiles. Validation of the data presented will be crucial to dissect the different regulatory mechanisms of PEs and their functional role in pluripotency, providing further insight into the gene regulatory control that characterizes the early stages of human development and pluripotency.

Bibliography

[1] Walther Flemming. Zellsubstanz, Kern und Zelltheilung. *F.C.W. Vogel Liepzig*, 1882.

[2] Ada L. Olins and Donald E. Olins. Spheroid chromatin units (v bodies). *Science*, 183(4122):330–332, 1974.

[3] C L F Woodcock, J P Safer, and J E Stanchfield. Structural repeating units in chromatin. *Experimental cell research.*, 97(1).

[4] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[5] Timothy J. Richmond and Curt A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, 2003.

[6] Michelle S. Ong, Timothy J. Richmond, and Curt A. Davey. DNA Stretching and Extreme Kinking in the Nucleosome Core. *Journal of Molecular Biology*, 368(4):1067–1074, 2007.

[7] Julianne L. Kitevski-Leblanc, Tairan Yuwen, Pamela N. Dyer, Johannes Rudolph, Karolin Luger, and Lewis E. Kay. Investigating the Dynamics of Destabilized Nucleosomes Using Methyl-TROSY NMR. *Journal of the American Chemical Society*, 140(14):4774–4777, 2018.

[8] James D. McGhee, Joanne M. Nickol, Gary Felsenfeld, and Donald C. Rau. Higher order structure of chromatin: Orientation of nucleosomes within the 30 nm chromatin solenoid is independent of species and spacer length. *Cell*, 33(3):831–841, 1983.

[9] D. Z. Staynov, S. Dunn, J. P. Baldwin, and C. Crane-Robinson. Nuclease digestion patterns as a criterion for nucleosome orientation in the higher order structure of chromatin. *FEBS Letters*, 157(2):311–315, 1983.

[10] C. L.F. Woodcock, L. L.Y. Frado, and J. B. Rattner. The higher-order structure of chromatin: Evidence for a helical ribbon arrangement. *Journal of Cell Biology*, 99(1 I):42–52, 1984.

[11] Benedetta Dorigo, Thomas Schalch, Alexandra Kulangara, Sylwia Duda, Rasmus R. Schroeder, and Timothy J. Richmond. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science*, 306(5701):1571–1573, 2004.

[12] Kazuhiro Maeshima, Saera Hihara, and Mikhail Eltsov. Chromatin structure: Does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, 22(3):291–297, 2010.

[13] Eden Fussner, Reagan W. Ching, and David P. Bazett-Jones. Living without 30nm chromatin fibers. *Trends in Biochemical Sciences*, 36(1):1–6, 2011.

[14] Margot P. Scheffer, Mikhail Eltsov, and Achilleas S. Frangakis. Evidence for short-range helical order in the 30-nm chromatin fibers of erythrocyte nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, 108(41):16992–16997, 2011.

[15] R. A. Horowitz, D. A. Agard, J. W. Sedat, and C. L. Woodcock. The three-dimensional architecture of chromatin in situ: Electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon. *Journal of Cell Biology*, 125(1):1–10, 1994.

[16] Job Dekker, Karsten Rippe, Martinjn Dekker, and Nancy Kleckner. Capturing Chromosome Conformation. *Science*, 295:1306–11, 2002.

[17] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Xu Han, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Yao Hui Huang, Willem-jan Welboren, Yuyuan Han, Hong-sain Ooi, N Pramila, Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-rataboul, Wing-kin Sung, Edison T Liu, and Chia-lin Wei. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2010.

[18] Erez Lieberman-aiden, Nynke L Van Berkum, Louise Williams, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, J Peter, Michael O Dorschner, Richard

Sandstrom, Bradley Bernstein, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, and A Leonid. NIH Public Access. 326(5950):289–293, 2010.

[19] Oluwatosin Oluwadare, Max Highsmith, and Jianlin Cheng. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biological Procedures Online*, 21(1):1–20, 2019.

[20] Tsung-han S Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and Oliver J Rando. Mapping nucleosome resolution chromosome folding in yeast by Micro-C. 162(1):108–119, 2016.

[21] Masae Ohno, Tadashi Ando, David G. Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs. *Cell*, 176(3):520–534, 2019.

[22] Horng D Ou, Sebastien Phan, Thomas J Deerinck, Andrea Thos, MarkH Ellisman, and Clodagh O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349), 2017.

[23] Carstem Carlberg and Ferdinand Molnar. Chromatin. In *Human Epigenetics: How Science Works*, pages 15–28. 2019.

[24] Nehmé Saksouk, Elisabeth Simboeck, and Jérôme Déjardin. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics and Chromatin*, 8(1):1–17, 2015.

[25] Robin C. Allshire and Hiten D. Madhani. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology*, 19(4):229–244, 2018.

[26] Paul B. Talbert and Steven Henikoff. Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly. *Trends in Genetics*, 34(8):587–599, 2018.

[27] René M. Marsano, E. Giordano, Giovanni Messina, and Patrizio Dimitri. A New Portrait of Constitutive Heterochromatin: Lessons from Drosophila melanogaster. *Trends in Genetics*, 35(9):615–631, 2019.

[28] Stefanie Rosa and Peter Shaw. Insights into chromatin structure and dynamics in plants. *Biology*, 2(4):1378–1410, 2013.

[29] R. Chen, R. Kang, X. G. Fan, and D. Tang. Release and activity of histone in diseases. *Cell Death and Disease*, 5(8):1–9, 2014.

[30] Michael Bulger. Hyperacetylated chromatin domains: Lessons from heterochromatin. *Journal of Biological Chemistry*, 280(23):21689–21692, 2005.

[31] Craig L. Peterson and Marc André Laniel. Histones and histone modifications. *Current biology : CB*, 14(14):546–551, 2004.

[32] Carla Sawan and Herceg Zdenko. Histone modifications and cancer.

[33] Taraswi Banerjee and Debabrata Chakravarti. A Peek into the Complex Realm of Histone Phosphorylation. *Molecular and Cellular Biology*, 31(24):4858–4873, 2011.

[34] Mariano Labrador and Victor G. Corces. Phosphorylation of histone H3 during transcriptional activation depends on promoter structure. *Genes and Development*, 17(1):43–48, 2003.

[35] Scott J. Nowak and Victor G. Corces. Phosphorylation of histone H3: A balancing act between chromosome condensation and transcriptional activation. *Trends in Genetics*, 20(4):214–220, 2004.

[36] S. C. Hodawadekar and R. Marmorstein. Chemistry of acetyl transfer by histone modifying enzymes: Structure, mechanism and implications for effector design. *Oncogene*, 26(37):5528–5540, 2007.

[37] Chang Huang, Mo Xu, and Bing Zhu. Epigenetic inheritance mediated by histone lysine methylation: Maintaining transcriptional states without the precise restoration of marks? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1609), 2013.

[38] Christopher Wood, Ambrosius Snijders, James Williamson, Colin Reynolds, John Baldwin, and Mark Dickman. Post-translational modifications of the linker histone variants and their association with cell mechanisms. *FEBS Journal*, 276(14):3685–3697, 2009.

[39] Anna Sadakierska-Chudy and Małgorzata Filip. A Comprehensive View of the Epigenetic Landscape. Part II: Histone Post-translational Modification, Nucleosome Level, and Chromatin Regulation by ncRNAs. *Neurotoxicity Research*, 27(2):172–197, 2015.

[40] James E Brownell, Jianxin Zhou, Tamara Ranalli, Ryuji Kobayashi, Diane G Edmondson, Sharon Y Roth, and C David Allis. Tetrahymena histone acetyltransferase

A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, 84(6):843–851, 1996.

[41] X. J. Yang and E. Seto. HATs and HDACs: From structure, function and regulation to novel strategies for therapy and prevention. *Oncogene*, 26(37):5310–5318, 2007.

[42] Yanshuo Han, Fadwa Tanios, Christian Reeps, Jian Zhang, Kristina Schwamborn, Hans Henning Eckstein, Alma Zernecke, and Jaroslav Pelisek. Histone acetylation and histone acetyltransferases show significant alterations in human abdominal aortic aneurysm. *Clinical Epigenetics*, 8(1):1–13, 2016.

[43] Loredana Verdone, Micaela Caserta, and Ernesto Di Mauro. Role of histone acetylation in the control of gene expression. *Biochemistry and Cell Biology*, 83(3):344–353, 2005.

[44] Robert E. Kingston and Geeta J. Narlikar. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes and Development*, 13(18):2339–2352, 1999.

[45] Andrew J. Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2005.

[46] Cyrus Martin and Yi Zhang. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology*, 6(11):838–849, 2005.

[47] Marion Lohrum, Hendrik G. Stunnenberg, and Colin Logie. The new frontier in cancer research: Deciphering cancer epigenetics. *International Journal of Biochemistry and Cell Biology*, 39(7-8):1450–1461, 2007.

[48] S. S. Ng, W. W. Yue, U. Oppermann, and R. J. Klose. Dynamic protein methylation in chromatin biology. *Cellular and Molecular Life Sciences*, 66(3):407–422, 2009.

[49] Tianyi Zhang, Sarah Cooper, and Neil Brockdorff. The interplay of histone modifications – writers that read. *EMBO reports*, 16(11):1467–1481, 2015.

[50] Shahin Ramazi, Abdollah Allahverdi, and Javad Zahiri. Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *Journal of Biosciences*, 45(1):3–9, 2020.

[51] Peter J. Rugg-Gunn, Brian J. Cox, Amy Ralston, and Janet Rossant. Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24):10783–10790, 2010.

[52] Orly L. Wapinski, Thomas Vierbuchen, Kun Qu, Qian Yi Lee, Soham Chanda, Daniel R. Fuentes, Paul G. Giresi, Yi Han Ng, Samuele Marro, Norma F. Neff, Daniela Drechsel, Ben Martynoga, Diogo S. Castro, Ashley E. Webb, Thomas C. Südhof, Anne Brunet, Francois Guillemot, Howard Y. Chang, and Marius Wernig. XHierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell*, 155(3):621, 2013.

[53] Woojin An. Combinatorial player for transcriptional regulation. In *Subcellular Biochemistry*, chapter Histone Ac, pages 355–374. 2007.

[54] Andrea Piunti and Ali Shilatifard. Epigenetic balance of gene expression by polycomb and compass families. *Science*, 352(6290), 2016.

[55] Sergi Aranda, Gloria Mas, and Luciano Di Croce. Regulation of gene transcription by Polycomb proteins. *Science Advances*, 1(11):1–15, 2015.

[56] Bernd Schuettengruber, Henri Marc Bourbon, Luciano Di Croce, and Giacomo Cavalli. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell*, 171(1):34–57, 2017.

[57] Enrique Blanco, Mar González-Ramírez, Anna Alcaine-Colet, Sergi Aranda, and Luciano Di Croce. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics*, 36(2):118–131, 2020.

[58] PH Lewis. New mutants report. 1947.

[59] E. B. Lewis. A gene complex controlling segmentation in Drosophila. *Nature*, 276(5688):565–570, 1978.

[60] Renato Paro. Imprinting a determined state into the chromatin of Drosophila. *Trends in Genetics*, 6(C):416–421, 1990.

[61] F. Pelegri and R. Lehmann. A role of polycomb group genes in the regulation of gap gene expression in Drosophila. *Genetics*, 136(4):1341–1353, 1994.

[62] Neil P Blackledge, Nathan R Rose, and Robert J Klose. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nature Reviews Molecular Cell Biology*, 16(11):643–649, 2015.

[63] Angela K. Robinson, Belinda Z. Leal, Linda V. Chadwell, Renjing Wang, Udayar Ilangovan, Yogeet Kaur, Sarah E. Junco, Virgil Schirf, Pawel A. Osmulski, Maria Gaczynska, Andrew P. Hinck, Borries Demeler, Donald G. McEwen, and Chongwoo A. Kim. The growth-suppressive function of the polycomb group protein polyhomeotic is mediated by polymerization of its sterile alpha motif (SAM) domain. *Journal of Biological Chemistry*, 287(12):8702–8713, 2012.

[64] Frank Wilkinson, Heather Pratt, and Micheal L Atchinson. PcG Recruitment by the YY1 REPO Domain Can Be Mediated by Yaf2. *Journal of Cell Biochemistry*, 109(3):478–486, 2010.

[65] Simon Hauri, Federico Comoglio, Makiko Seimiya, Moritz Gerstung, Timo Glatter, Klaus Hansen, Ruedi Aebersold, Renato Paro, Matthias Gstaiger, and Christian Beisel. A High-Density Map for Navigating the Human Polycomb Complexome. *Cell Reports*, 17(2):583–595, 2016.

[66] Artyom A. Alekseyenko, Andrey A. Gorchakov, Peter V. Kharchenko, and Mitzi I. Kuroda. Reciprocal interactions of human C10orf12 and C17orf96 with PRC2 revealed by BioTAP-XL cross-linking and affinity purification. *Proceedings of the National Academy of Sciences of the United States of America*, 111(7):2488–2493, 2014.

[67] Susan L. Kloet, Matthew M. Makowski, H. Irem Baymaz, Lisa Van Voorthuijsen, Ino D. Karemaker, Alexandra Santanach, Pascal W.T.C. Jansen, Luciano Di Croce, and Michiel Vermeulen. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nature Structural and Molecular Biology*, 23(7):682–690, 2016.

[68] Ballare Cecilia, Martin Lange, Audrone Lapinaite, Gloria Mas Martin, Lluis Morey, Gloria Pascual, Robert Liefke, Bernd Simon, Yang Shi, Or Gozani, Teresa Carlomagno, Salvador Aznar Benitah, and Luciano Di Croce. Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nature Structural and Molecular Biology*, 19(12):1257–1265, 2012.

[69] Gerard L. Brien, Guillermo Gambero, David J. O'Connell, Emilia Jerman, Siobhán A. Turner, Chris M. Egan, Eiseart J. Dunne, Maike C. Jurgens, Kieran Wynne, Lianhua Piao, Amanda J. Lohan, Neil Ferguson, Xiaobing Shi, Krishna M. Sinha, Brendan J. Loftus, Gerard Cagney, and Adrian P. Bracken. Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nature Structural and Molecular Biology*, 19(12):1273–1281, 2012.

[70] Haojie Li, Robert Liefke, Junyi Jiang, Jesse Vigoda Kurland, Wei Tian, Pujuan Deng, Weidi Zhang, Qian He, Dinshaw J. Patel, Martha L. Bulyk, Yang Shi, and Zhanxin Wang. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature*, 549(7671):287–291, 2017.

[71] Matteo Perino, Guido Van Mierlo, Ino D. Karemaker, Siebe Van Genesen, Michiel Vermeulen, Hendrik Marks, Simon J. Van Heeringen, and Gert Jan C. Veenstra. MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. *Nature Genetics*, 50(7):1002–1010, 2018.

[72] Ling Cai, Scott B. Rothbart, Rui Lu, Bowen Xu, Wei Yi Chen, Ashutosh Tripathy, Shira Rockowitz, Deyou Zheng, Dinshaw J. Patel, C. David Allis, Brian D. Strahl, Jikui Song, and Gang Greg Wang. An H3K36 Methylation-Engaging Tudor Motif of Polycomb-like Proteins Mediates PRC2 Complex Targeting. *Molecular Cell*, 49(3):571–582, 2013.

[73] Malte Beringer, Paola Pisano, Valerio Di Carlo, Enrique Blanco, Paul Chammas, Pedro Vizán, Arantxa Gutiérrez, Sergi Aranda, Bernhard Payer, Michael Wierer, and Luciano Di Croce. EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. *Molecular Cell*, 64(4):645–658, 2016.

[74] Robert Liefke, Violetta Karwacki-Neisius, and Yang Shi. EPOP interacts with Elongin BC and USP7 to modulate the chromatin landscape. *Molecular Cell*, 64(4):659–672, 2016.

[75] Eric Conway, Emilia Jerman, Evan Healy, Shinsuke Ito, Daniel Holoch, Giorgio Oliviero, Orla Deevy, Eleanor Glancy, Darren J. Fitzpatrick, Marlena Mucha, Ariane Watson, Alan M. Rice, Paul Chammas, Christine Huang, Indigo Pratt-Kelly, Yoko Koseki, Manabu Nakayama, Tomoyuki Ishikura, Gundula Streubel, Kieran

Wynne, Karsten Hokamp, Aoife McLysaght, Claudio Ciferri, Luciano Di Croce, Gerard Cagney, Raphaël Margueron, Haruhiko Koseki, and Adrian P. Bracken. A Family of Vertebrate-Specific Polycombs Encoded by the LCOR/LCORL Genes Balance PRC2 Subtype Activities. *Molecular Cell*, 70(3):408–421, 2018.

[76] Jamy C Peng, Anton Valouev, Tomek Swigut, Junmei Zhang, Yingming Zhao, Arend Sidow, and Joanna Wysocka. Jarid2/Jumonji Coordinates Control of PRC2 Enzymatic Activity and Target Gene Occupancy in Pluripotent Cells. *Cell*, 139(7):1290–1302, 2009.

[77] Gang Li, Raphael Margueron, Manching Ku, Pierre Chambon, Bradley E. Bernstein, and Danny Reinberg. Jarid2 and PRC2, partners in regulating gene expression. *Genes and Development*, 24(4):368–380, 2010.

[78] Reinhard Kalb, Sebastian Latwiel, H. Irem Baymaz, Pascal W.T.C. Jansen, Christoph W. Müller, Michiel Vermeulen, and Jürg Müller. Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nature Structural and Molecular Biology*, 21(6):569–571, 2014.

[79] Sarah Cooper, Anne Grijzenhout, Elizabeth Underwood, Katia Ancelin, Tianyi Zhang, Tatyana B. Nesterova, Burcu Anil-Kirmizitas, Andrew Bassett, Susanne M. Kooistra, Karl Agger, Kristian Helin, Edith Heard, and Neil Brockdorff. Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. *Nature Communications*, 7:1–8, 2016.

[80] Xueyin Wang, Richard D Paucek, Anne R Gooding, Zachary Z Brown, Eva J Ge, Tom W Muir, and Thomas R Cech. Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nature Structural and Molecular Biology*, 24(12):1028–1038, 2017.

[81] Bernd Schuettengruber, Daniel Chourrout, Michel Vervoort, Benjamin Leblanc, and Giacomo Cavalli. Genome Regulation by Polycomb and Trithorax Proteins. *Cell*, 128(4):735–745, 2007.

[82] Lluis Morey, Gloria Pascual, Luca Cozzuto, Guglielmo Roma, Anton Wutz, Salvador Aznar Benitah, and Luciano Di Croce. Nonoverlapping functions of the polycomb group Cbx family of proteins in embryonic stem cells. *Cell Stem Cell*, 10(1):47–62, 2012.

[83] Ana O'Loghlen, Ana M. Muñoz-Cabello, Alexandre Gaspar-Maia, Hsan Au Wu, Ana Banito, Natalia Kunowska, Tomas Racek, Helen N. Pemberton, Patrizia Beolchi, Fabrice Lavial, Osamu Masui, Michiel Vermeulen, Thomas Carroll, Johannes Graumann, Edith Heard, Niall Dillon, Veronique Azuara, Ambrosius P. Snijders, Gordon Peters, Emily Bernstein, and Jesus Gil. MicroRNA regulation of Cbx7 mediates a switch of polycomb orthologs during ESC differentiation. *Cell Stem Cell*, 10(1):33–46, 2012.

[84] Jeffrey Simon, Anne Chiang, Welcome Bender, Mary Jane Shimell, and Micheal O'Connor. Elements of the Drosophila Bithorax Complex That Mediate Repression by Polycomb Group Products. *Developmental Biology*, 158(1):131–144, 1993.

[85] J. Muller and M. Bienz. Long range repression conferring boundaries of Ultrabithorax expression in the Drosophila embryo. *EMBO Journal*, 10(11):3147–3155, 1991.

[86] M. O. Fauvarque and J. M. Dura. polyhomeotic Regulatory sequences induce developmental regulator-dependent variegation and targeted P-element insertions in Drosophila. *Genes and Development*, 7(8):1508–1520, 1993.

[87] Marianne Entrevan, Bernd Schuettengruber, and Giacomo Cavalli. Regulation of Genome Architecture and Function by Polycomb Proteins. *Trends in Cell Biology*, 26(7):511–525, 2016.

[88] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando J. Miller, and Donald Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99, 1985.

[89] Sally H. Cross, Jillian A. Charlton, Xinsheng Nan, and Adrian P. Bird. Purification of CpG islands using a methylated DNA binding column. *Nature Genetics*, 6(3):236–244, 1994.

[90] Serge Saxonov, Paul Berg, and Douglas L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–1417, 2006.

[91] Eric M. Mendenhall, Richard P. Koche, Thanh Truong, Vicky W. Zhou, Biju Issac, Andrew S. Chi, Manching Ku, and Bradley E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genetics*, 6(12):1–10, 2010.

[92] Eva Madi Riising, Itys Comet, Benjamin Leblanc, Xudong Wu, Jens Vilstrup Johansen, and Kristian Helin. Gene silencing triggers polycomb repressive complex 2 recruitment to CpG Islands genome wide. *Molecular Cell*, 55(3):347–360, 2014.

[93] Anca M. Farcas, Neil P. Blackledge, Ian Sudbery, Hannah K. Long, Joanna F. McGouran, Nathan R. Rose, Sheena Lee, David Sims, Andrea Cerase, Thomas W. Sheahan, Haruhiko Koseki, Neil Brockdorff, Chris P. Ponting, Benedikt M. Kessler, and Robert J. Klose. KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. *eLife*, 2012(1):1–26, 2012.

[94] Robert J. Klose, Sarah Cooper, Anca M. Farcas, Neil P. Blackledge, and Neil Brockdorff. Chromatin Sampling-An Emerging Perspective on Targeting Polycomb Repressor Proteins. *PLoS Genetics*, 9(8), 2013.

[95] Zhaohui Shao, Florian Raible, Ramin Mollaaghababa, Jeffrey R. Guyon, Chao Ting Wu, Welcome Bender, and Robert E. Kingston. Stabilization of chromatin structure by PRC1, a polycomb complex. *Cell*, 98(1):37–46, 1999.

[96] Feng Tie, Rakhee Banerjee, Chen Fu, Carl A. Stratton, Ming Fang, and Peter J. Harte. Polycomb inhibits histone acetylation by CBP by binding directly to its catalytic domain. *Proceedings of the National Academy of Sciences of the United States of America*, 113(6):E744–E753, 2016.

[97] Gaetano I. Dellino, Yuri B. Schwartz, Gabriella Farkas, Donna McCabe, Sarah C.R. Elgin, and Vincenzo Pirrotta. Polycomb silencing blocks transcription initiation. *Molecular Cell*, 13(6):887–893, 2004.

[98] Wenlai Zhou, Ping Zhu, Jianxun Wang, Gabriel Pascual, Kenneth A Ohgi, Jean Lozach, Christopher K Glass, and Micheal G Rosenfeld. Histone H2A Monoubiquitination Represses Transcription by Inhibiting RNA Polymerase II Transcriptional Elongation. *Molecular Cell*, 29(1):60–80, 2008.

[99] Takeya Nakagawa, Takuya Kajitani, Shinji Togo, Norio Masuko, Hideki Ohdan, Yoshitaka Hishikawa, Takehiko Koji, Toshifumi Matsuyama, Tsuyoshi Ikura, Masami Muramatsu, and Takashi Ito. Deubiquitylation of histone H2A activates transcriptional initiation via trans-histone cross-talk with H3K4 di- and trimethylation. *Genes and Development*, 22(1):37–49, 2008.

[100] Andrea Cerase, Daniel Smeets, Y. Amy Tang, Michal Gdula, Felix Kraus, Mikhail Spivakov, Benoit Moindrot, Marion Leleu, Anna Tattermusch, Justin Demmerle, Tatyana B. Nesterova, Catherine Green, Arie P. Otte, Lothar Schermelleh, and Neil Brockdorff. Spatial separation of Xist RNA and polycomb proteins revealed by superresolution microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 111(6):2235–2240, 2014.

[101] Catherine Cifuente-Rojas, Alfredo J Hernandez, Kavitha Sarma, and Jeannie T. Lee. Regulatory interactions between RNA and Polycomb Repressive Complex 2. *Molecular Cell*, 55(2):171–185, 2014.

[102] Veronika A. Herzog, Adelheid Lempradl, Johanna Trupke, Helena Okulski, Christina Altmutter, Frank Ruge, Bernd Boidol, Stefan Kubicek, Gerald Schmauss, Karin Aumayr, Marius Ruf, Andrew Pospisilik, Andrew Dimond, Hasene Basak Senergin, Marcus L. Vargas, Jeffrey A. Simon, and Leonie Ringrose. A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. *Nature Genetics*, 46(9):973–981, 2014.

[103] Manuel Beltran, Christopher M. Yates, Lenka Skalska, Marcus Dawson, Filipa P. Reis, Keijo Viiri, Cynthia L. Fisher, Christopher R. Sibley, Benjamin M. Foster, Till Bartke, Jernej Ule, and Richard G. Jenner. The interaction of PRC2 with RNA or chromatin s mutually antagonistic. *Genome Research*, 26(7):896–907, 2016.

[104] Manuela Portoso, Roberta Ragazzini, Brencic Ziva, Arianna Moiani, Audrey Michaud, Ivaylo Vassilev, Michel Wassef, Nicolas Servant, Bruno Sargueil, and Raphaël Margueron. PRC 2 is dispensable for HOTAIR -mediated transcriptional repression. *The EMBO Journal*, 36(8):981–994, 2017.

[105] Kathrin Plath, Jia Fang, Susanna K. Mlynarczyk-Evans, Ru Cao, Kathleen A. Worringer, Hengbin Wang, Cecile C. De la Cruz, Arie P. Otte, Barbara Panning, and Yi Zhang. Role of histone H3 lysine 27 methylation in X inactivation. *Science*, 300(5616):131–135, 2013.

[106] Mafalda Almeida, Joseph S. Bowness, and Neil Brockdorff. The many faces of Polycomb regulation by RNA. *Current Opinion in Genetics and Development*, 61:53–61, 2020.

[107] Haining Zhou, Chad B Stein, Tiasha A Shafiq, Gergana Shipkovenska, Marian Kalocsay, Joao A Paulo, Jiuchun Zhang, Zhenhua Luo, Steven P Gygi, Karen Adelman, and Danesh Moazed. Rixosomal RNA degradation contributes to silencing of Polycomb target genes. *Nature*, 2022.

[108] Frédéric Bantignies, Virginie Roure, Itys Comet, Benjamin Leblanc, Bernd Schuettengruber, Jérôme Bonnet, Vanessa Tixier, André Mas, and Giacomo Cavalli. Polycomb-dependent regulatory contacts between distant hox loci in drosophila. *Cell*, 144(2):214–226, 2011.

[109] Thierry Cheutin and Giacomo Cavalli. Polycomb silencing: From linear chromatin domains to 3D chromosome folding. *Current Opinion in Genetics and Development*, 25(1):30–37, 2014.

[110] Ajazul H. Wani, Alistair N. Boettiger, Patrick Schorderet, Ayla Ergun, Christine Munger, Ruslan I. Sadreyev, Xiaowei Zhuang, Robert E. Kingston, and Nicole J. Francis. Chromatin topology is coupled to Polycomb group protein subnuclear organization. *Nature Communications*, 7, 2016.

[111] Tomas Pachano, Giuliano Crispatzu, and Alvaro Rada-Iglesias. Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Briefings in Functional Genomics*, 18(6):358–366, 2019.

[112] Jean Philippe Fortin and Kasper D. Hansen. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, 16(1):1–23, 2015.

[113] Tapan Kumar Mohanta, Awdhesh Kumar Mishra, and Ahmed Al-Harrasi. The 3d genome: From structure to function. *International Journal of Molecular Sciences*, 22(21):1–31, 2021.

[114] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, Annelies De Klein, Lodewyk Wessels, Wouter De Laat, and Bas Van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, 2008.

[115] Brian Burke and Colin L. Stewart. The nuclear lamins: Flexibility in function. *Nature Reviews Molecular Cell Biology*, 14(1):13–24, 2013.

[116] Ana Pombo and Niall Dillon. Three-dimensional genome architecture: Players and mechanisms. *Nature Reviews Molecular Cell Biology*, 16(4):245–257, 2015.

[117] Wouter Meuleman, Daan Peric-Hupkes, Jop Kind, Jean Bernard Beaudry, Ludo Pagie, Manolis Kellis, Marcel Reinders, Lodewyk Wessels, and Bas Van Steensel. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Research*, 23(2):270–280, 2013.

[118] David R. Buckler, Yuchen Zhou, and Ann M. Stock. Evidence of intradomain and interdomain flexibility in an OmpR/PhoB Homolog from Thermotoga maritima. *Structure*, 10(2):153–164, 2002.

[119] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[120] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, 9(1), 2018.

[121] Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4), 2019.

[122] Judita Richterova, Barbora Huraiova, and Juraj Gregan. Genome Organization: Cohesin on the Move. *Molecular Cell*, 66(4):444–445, 2017.

[123] Georg A. Busslinger, Roman R. Stocsits, Petra Van Der Lelij, Elin Axelsson, Antonio Tedeschi, Niels Galjart, and Jan Michael Peters. Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, 544(7651):503–507, 2017.

[124] Haifeng Wang, Mengting Han, and Lei S. Qi. Engineering 3D genome organization. *Nature Reviews Genetics*, 22(6):343–360, 2021.

[125] Maria A. Ferraiuolo, Mathieu Rousseau, Carol Miyamoto, Solomon Shenker, Xue Qing David Wang, Michelle Nadler, Mathieu Blanchette, and Josée Dostie. The

three-dimensional architecture of Hox cluster silencing. *Nucleic Acids Research*, 38(21):7472–7484, 2010.

[126] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3):557–572, 2017.

[127] Maxence Vieux-Rochas, Pierre J. Fabre, Marion Leleu, Denis Duboule, and Daan Noordermeer. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15):4672–4677, 2015.

[128] James D.P. Rhodes, Angelika Feldmann, Benjamín Hernández-Rodríguez, Noelia Díaz, Jill M. Brown, Nadezda A. Fursova, Neil P. Blackledge, Praveen Prathapan, Paula Dobrinic, Miles K. Huseyin, Aleksander Szczurek, Kai Kruse, Kim A. Nasmyth, Veronica J. Buckle, Juan M. Vaquerizas, and Robert J. Klose. Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells. *Cell Reports*, 30(3):820–835, 2020.

[129] Gozde Kar, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Kedar Nath Natarajan, Elena Torlai Triglia, Borbala Mifsud, Sarah Elderkin, John C. Marioni, Ana Pombo, and Sarah A. Teichmann. Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *Nature Communications*, 8(1):1–13, 2017.

[130] Vincent Loubiere, Anne Marie Martinez, and Giacomo Cavalli. Cell Fate and Developmental Regulation Dynamics by Polycomb Proteins and 3D Genome Architecture. *BioEssays*, 41(3):1–15, 2019.

[131] T Boveri. Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität. *Engelmann*, 1909.

[132] E. Passarge. Emil Heitz and the concept of heterochromatin: Longitudinal chromosome differentiation was recognized fifty years ago. *American Journal of Human Genetics*, 31(2):106–115, 1979.

[133] T Cremer, C Cremer, T Schneider, H Baumann, L Hens, and M Krisch-Volders. Analysis of chromosome positions in the interphase nucleus of Chinese hamster

cells by laser-UV-microirradiation experiments. *Human Genetics*, 62(3):201–209, 1982.

[134] L Manuelidis. Individual interphase chromosome domains revealed by in situ hybridization. *Human Genetics*, 71(4):288–293, 1985.

[135] Margit Schardin, T. Cremer, H. D. Hager, and M. Lang. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Human Genetics*, 71(4):281–287, 1985.

[136] Katherine E. Cullen, Michael P. Kladde, and Mark A. Seyfred. Interaction between transcription regulatory regions of prolactin chromatin. *Science*, 261(5118):203–206, 1993.

[137] Rieke Kempfer and Ana Pombo. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, 21(4):207–226, 2020.

[138] Claire Marchal, Jiao Sima, and David M. Gilbert. Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology*, 20(12):721–737, 2019.

[139] Marco Di Stefano, Francesca Di Giovanni, Vasilisa Pozharskaia, Mercè Gomar-Alba, Davide Baù, Lucas B. Carey, Marc A. Marti-Renom, and Manuel Mendoza. Impact of Chromosome Fusions on 3D Genome. 214(March):651–667, 2020.

[140] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38(11):1348–1354, 2006.

[141] Mathieu Rousseau, Jennifer L. Crutchley, Hisashi Miura, Matthew Suderman, Mathieu Blanchette, and Josée Dostie. Hox in motion: Tracking HoxA cluster conformation during differentiation. *Nucleic Acids Research*, 42(3):1524–1540, 2014.

[142] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. Van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[143] Ivana Jerkovic´ and Giacomo Cavalli. Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, 22(8):511–528, 2021.

[144] Vanja Haberle and Alexander Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, 2018.

[145] RG Roeder and WJ Rutter. Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature*, 224:1969, 1969.

[146] Andre Sentenac. Eukaryotic RNA polymerase. *Critical Reviews in Biochemistry and Molecular Biology*, 18(1):31–90, 1985.

[147] Nicholas J. Fuda, M. Behfar Ardehali, and John T. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192, 2009.

[148] Dustin E. Schones, Kairong Cui, Suresh Cuddapah, Tae Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5):887–898, 2008.

[149] Aimée M. Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes and Development*, 25(10):1010–1022, 2011.

[150] Ferenc Müller and Làszlò Tora. Chromatin and DNA sequences in defining promoters for transcription initiation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1839(3):118–128, 2014.

[151] Long Vo Ngoc, George A. Kassavetis, and James T. Kadonaga. The RNA polymerase II core promoter in Drosophila. *Genetics*, 212(1):13–24, 2019.

[152] Paul B. Talbert, Michael P. Meers, and Steven Henikoff. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nature Reviews Genetics*, 20(5):283–297, 2019.

[153] D Reinberg, G Orphanides, R Ebright, S Akoulitchev, J Carcamo, H Cho, P Cortes, R Drapkin, O Flores, I Ha, JA Inastroza, S Kim, TK Kim, P Kumar, T Lagrange, G Leroy, H Lu, DM Ma, E Maldonando, A Merino, F Mermelstein, I Olave, M Sheldon, R Shiekhattar, N Stone, X Sun, L Weis, K Yeung, and L Zawel. The RNA

Polymerase II General Transcription Factors: Past, Present, and Future. *Cold Spring Harb Symp Quant Biol*, 63:83–105, 1998.

[154] Ingrid Grummt. Life on a planet of its own: Regulation of RNA polymerase I transcription in the nucleolus. *Genes and Development*, 17(14):1691–1702, 2003.

[155] Laura Schramm and Nouria Hernandez. Recruitment of RNA polymerase III to its target promoters. *Genes and Development*, 16(20):2593–2620, 2002.

[156] E. Peter Geiduschek and George A. Kassavetis. The RNA polymerase III transcription apparatus. *Journal of Molecular Biology*, 310(1):1–26, 2001.

[157] André Sentenac and Michel Riva. Odd RNA polymerases or the A(B)C of eukaryotic transcription. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829(3-4):251–257, 2013.

[158] Dirk Kostrewa, Mirijam E. Zeller, Karim Jean Armache, Martin Seizl, Kristin Leike, Michael Thomm, and Patrick Cramer. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature*, 462(7271):323–330, 2009.

[159] Hung Ta Chen and Steven Hahn. Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: A model for the structure of the PIC. *Cell*, 119(2):169–180, 2004.

[160] David A. Bushnell, Kenneth D. Westover, Ralph E. Davis, and Roger D. Kornberg. Structural Basis of Transcription: An RNA Polymerase II-TFIIB Cocrystal at 4.5 Angstroms. *Science*, 303(5660):983–988, 2004.

[161] Sarah Sainsbury, Jürgen Niesser, and Patrick Cramer. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature*, 493(7432):437–440, 2013.

[162] Jean Marc Egly and Frédéric Coin. A history of TFIIH: Two decades of molecular biology on a pivotal transcription/repair factor. *DNA Repair*, 10(7):714–721, 2011.

[163] Tae Kyung Kim, Richard H. Ebright, and Danny Reinberg. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science*, 288(5470):1418–1421, 2000.

[164] Roger D. Kornberg. Mediator and the mechanism of transcriptional activation. *Trends in Biochemical Sciences*, 30(5):235–239, 2005.

[165] Christian Dienemann, Björn Schwalb, Sandra Schilbach, and Patrick Cramer. Promoter Distortion and Opening in the RNA Polymerase II Cleft. *Molecular Cell*, 73(1):97–106, 2019.

[166] James Fishburn, Eric Galburt, and Steven Hahn. Transcription start site scanning and the requirement for ATP during transcription initiation by RNA polymerase II. *Journal of Biological Chemistry*, 291(25):13040–13047, 2016.

[167] S. Schilbach, M. Hantsche, D. Tegunov, C. Dienemann, C. Wigge, H. Urlaub, and P. Cramer. Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature*, 551(7679):204–209, 2017.

[168] Leighton Core and Karen Adelman. Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes and Development*, 33(15-16):960–982, 2019.

[169] John T. Lis. A 50 year history of technologies that drove discovery in eukaryotic transcription regulation. *Nature Structural and Molecular Biology*, 26(9):777–782, 2019.

[170] Donal S. Luse, Mrutyunjaya Parida, Benjamin M. Spector, Kyle A. Nilson, and David H. Price. A unified view of the sequence and functional organization of the human RNA polymerase II promoter. *Nucleic Acids Research*, 48(14):7767–7785, 2020.

[171] D S Gilmour and J T Lis. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Molecular and Cellular Biology*, 6(11):3984–3989, 1986.

[172] Ann E. Rougvie and John T. Lis. The RNA polymerase II molecule at the 5 end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6):795–804, 1988.

[173] Bjoern Gaertner and Julia Zeitlinger. RNA polymerase II pausing during development. *Development*, 141(6):1179–1183, 2014.

[174] Andreas Mayer, Heather M. Landry, and L. Stirling Churchman. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Current Opinion in Cell Biology*, 46:72–80, 2017.

[175] Kevin M. Harlen and L. Stirling Churchman. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology*, 18(4):263–273, 2017.

[176] Seychelle M. Vos, Lucas Farnung, Henning Urlaub, and Patrick Cramer. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature*, 560(7720):601–606, 2018.

[177] Yuki Yamaguchi, Hirotaka Shibata, and Hiroshi Handa. Transcription elongation factors DSIF and NELF: Promoter-proximal pausing and beyond. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829(1):98–104, 2013.

[178] Allison C. Schier and Dylan J. Taatjes. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes and Development*, 34(7-8):465–488, 2020.

[179] João D. Dias, Tiago Rito, Elena Torlai Triglia, Alexander Kukalev, Carmelo Ferrai, Mita Chotalia, Emily Brookes, Hiroshi Kimura, and Ana Pombo. Methylation of RNA polymerase II non-consensus Lysine residues marks early transcription in mammalian cells. *eLife*, 4(DECEMBER2015):1–30, 2015.

[180] Corinne N. Simonti, Katherine S. Pollard, Sebastian Schröder, Daniel He, Benoit G. Bruneau, Melanie Ott, and John A. Capra. Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. *BMC Evolutionary Biology*, 15(1):1–12, 2015.

[181] Julie Soutourina. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, 19(4):262–274, 2018.

[182] Carmelo Ferrai, Elena Torlai Triglia, Jessica R Risner-Janiczek, Tiago Rito, Owen JL Rackham, Inês Santiago, Alexander Kukalev, Mario Nicodemi, Altuna Akalin, Meng Li, Mark A Ungless, and Ana Pombo. RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation . *Molecular Systems Biology*, 13(10):946, 2017.

[183] Julie K. Stock, Sara Giadrossi, Miguel Casanova, Emily Brookes, Miguel Vidal, Haruhiko Koseki, Neil Brockdorff, Amanda G. Fisher, and Ana Pombo. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nature Cell Biology*, 9(12):1428–1435, 2007.

[184] Emily Brookes and Ana Pombo. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Reports*, 10(11):1213–1219, 2009.

[185] Wee Wei Tee, Steven S. Shen, Ozgur Oksuz, Varun Narendra, and Danny Reinberg. Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. *Cell*, 156(4):678–690, 2014.

[186] Chun Ma, Violetta Karwacki-Neisius, Haoran Tang, Wenjing Li, Zhennan Shi, Haolin Hu, Wenqui Xu, Zehentian Wang, Lingchun Kong, Ruitu Lv, Zheng Fan, Wenhao Zhou, Pengyuan Yang, Feizhen Wu, Jianbo Diao, Li Tan, Yujiang Geno Shi, Fei Lan, and Yang Shi. Nono, a bivalent domain factor, regulates Erk signaling and mouse embryonic stem cell pluripotency. *Cell Reports*, 17(4):997–1007, 2016.

[187] Andrew Field and Karen Adelman. Evaluating Enhancer Function and Transcription. *Annual Review of Biochemistry*, 89:213–234, 2020.

[188] Helen Ray-Jones and Mikhail Spivakov. *Transcriptional enhancers and their communication with gene promoters*, volume 78. Springer International Publishing, 2021.

[189] MArio Capecchi. High Efficiency Transformation by Direct Microinjection of DNA into Cultured Mammalian Cells. *Cell*, 22:479–488, 1980.

[190] R. Grosschedl and M. L. Birnstiel. Spacer DNA sequences upstream of the TATAAATA sequence are essential for promotion of H2A histone gene transcription in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 77(12 II):7102–7106, 1980.

[191] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 PART 1):299–308, 1981.

[192] P Moreau, R Hen, B Wasylyk, R Everett, MP Gaub, and P Chambon. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Research*, 9(22), 1981.

[193] Christophe Benoist and Pierre Chambon. In vivo sequence requirements of the SV40 early promoter region. *Nature*, 290(5804):304–310, 1981.

[194] P. Gruss, R. Dhar, and G. Khoury. Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 78(2 II):943–947, 1981.

[195] K. Struhl. Genetic properties and chromatin structure of the yeast gal regulatory element: An enhancer-like sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 81(24 I):7865–7869, 1984.

[196] B Shepherd, MJ Garabedian, MC Hung, and PC Wensink. Developmental Control of Drosophila Yolk Protein 1 Gene by cis-acting DNA Elements. *Cold Spring Harb Symp Quant Biol*, 50:521–526, 1985.

[197] S D Gillies, S L Morrison, V T Oi, and S Tonegawa. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, 33(3):717–728, 1983.

[198] Julian Banerji, Laura Olson, and Walter Schaffner. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3):729–740, 1983.

[199] Mark Mercola, Xiao Fan Wang, Jory Olsen, and Kathryn Calame. Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus. *Science*, 221(4611):663–665, 1983.

[200] Yin Shen, Feng Yue, David F. Mc Cleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Bing Ren, and Victor V. Lobanenkov. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[201] Jiang Zhu, Mazhar Adli, James Y. Zou, Griet Verstappen, Michael Coyne, Xiaolan Zhang, Timothy Durham, Mohammad Miri, Vikram Deshpande, Philip L. De Jager, David A. Bennett, Joseph A. Houmard, Deborah M. Muoio, Tamer T. Onder, Ray Camahort, Chad A. Cowan, Alexander Meissner, Charles B. Epstein, Noam Shoresh, and Bradley E. Bernstein. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–654, 2013.

[202] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, and Jennifer Harrow. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[203] Sven Heinz, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass.

The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3):144–154, 2015.

[204] François Spitz and Eileen E.M. Furlong. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

[205] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, 2016.

[206] Kenneth S. Zaret. Pioneering the chromatin landscape. *Nature Genetics*, 50(2):167–169, 2018.

[207] Meilin Fernandez Garcia, Cedric D. Moore, Katharine N. Schulz, Oscar Alberto, Greg Donague, Melissa M. Harrison, Heng Zhu, and Kenneth S. Zaret. Structural Features of Transcription Factors Associating with Nucleosome Binding. *Molecular Cell*, 75(5):921–932, 2019.

[208] Hsiao Lan Liang, Chung Yi Nien, Hsiao Yun Liu, Mark M. Metzstein, Nikolai Kirov, and Christine Rushlow. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, 456(7220):400–403, 2008.

[209] Melissa M. Harrison, Xiao Yong Li, Tommy Kaplan, Michael R. Botchan, and Michael B. Eisen. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genetics*, 7(10), 2011.

[210] Katharine N. Schulz, Eliana R. Bondra, Arbel Moshe, Jacqueline E. Villalta, Jason D. Lieb, Tommy Kaplan, Daniel J. McKay, and Melissa M. Harrison. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. *Genome Research*, 25(11):1715–1726, 2015.

[211] Laurie A. Boyer, Ihn Lee Tong, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Matthew G. Guenther, Roshan M. Kumar, Heather L. Murray, Richard G. Jenner, David K. Gifford, Douglas A. Melton, Rudolf Jaenisch, and Richard A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005.

[212] Xinyang Yu and Michael J. Buck. Pioneer factors and their in vitro identification methods. *Molecular Genetics and Genomics*, 295(4):825–835, 2020.

[213] Kenneth S. Zaret and Jason S. Carroll. Pioneer transcription factors: Establishing competence for gene expression. *Genes and Development*, 25(21):2227–2241, 2011.

[214] Robert G. Roeder. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Letters*, 579(4 SPEC. ISS.):909–915, 2005.

[215] Vikki M. Weake and Jerry L. Workman. Inducible gene expression: Diverse regulatory mechanisms. *Nature Reviews Genetics*, 11(6):426–437, 2010.

[216] Jian Sun, Yilin Zhao, Rebecca McGreal, Yamit Cohen-Tayar, Shira Rockowitz, Carola Wilczek, Ruth Ashery-Padan, David Shechter, Deyou Zheng, and Ales Cvekl. Pax6 associates with H3K4-specific histone methyltransferases Mll1, Mll2, and Set1a and regulates H3K4 methylation at promoters and enhancers. *Epigenetics and Chromatin*, 9(1):1–18, 2016.

[217] Kamila M. Jozwik, Igor Chernukhin, Aurelien A. Serandour, Sankari Nagarajan, and Jason S. Carroll. FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3. *Cell Reports*, 17(10):2715–2723, 2016.

[218] Kihyun Lee, Hyunwoo Cho, Robert W. Rickert, Qing V. Li, Julian Pulecio, Christina S. Leslie, and Danwei Huangfu. FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. *Cell Reports*, 28(2):382–393, 2019.

[219] Warren A. Whyte, David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, 2013.

[220] Ha Youn Shin, Michaela Willi, Kyung Hyun Yoo, Xianke Zeng, Chaochen Wang, Gil Metser, and Lothar Hennighausen. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature Genetics*, 48(8):904–911, 2016.

[221] Deborah Hay, Jim R. Hughes, Christian Babbs, James O.J. Davies, Bryony J. Graham, Lars L.P. Hanssen, Mira T. Kassouf, A. Marieke Oudelaar, Jacqueline A. Sharpe, Maria C. Suciu, Jelena Telenius, Ruth Williams, Christina Rode, Pik Shan Li, Len A.

Pennacchio, Jacqueline A. Sloane-Stanley, Helena Ayyub, Sue Butler, Tatjana Sauka-Spengler, Richard J. Gibbons, Andrew J.H. Smith, William G. Wood, and Douglas R. Higgs. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903, 2016.

[222] Noah Dukler, Brad Gulko, Yi Fei Huang, and Adam Siepel. Is a super-enhancer greater than the sum of its parts? *Nature Genetics*, 49(1):2–4, 2017.

[223] Dimitris Thanos and Tom Maniatis. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100, 1995.

[224] Daniel Panne, Tom Maniatis, and Stephen C. Harrison. An Atomic Model of the Interferon-β Enhanceosome. *Cell*, 129(6):1111–1123, 2007.

[225] Meghana M. Kulkarni and David N. Arnosti. Information display by transcriptional enhancers. *Development*, 130(26):6569–6575, 2003.

[226] David N. Arnosti and Meghana M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, 2005.

[227] Sepand Rastegar, Isabell Hess, Thomas Dickmeis, Jean Christophe Nicod, Raymond Ertzer, Yavor Hadzhiev, Wolf Gerolf Thies, Gerd Scherer, and Uwe Strähle. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Developmental Biology*, 318(2):366–377, 2008.

[228] Robin P. Smith, Leila Taher, Rupali P. Patwardhan, Mee J. Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45(9):1021–1028, 2013.

[229] Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, E. Hilary Gustafson, Ewan Birney, and Eileen E.M. Furlong. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, 148(3):473–486, 2012.

[230] Laura A. Lettice, Simon J.H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, B. A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14):1725–1735, 2003.

[231] Matthew J. Blow, David J. McCulley, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, James Bristow, Bing Ren, Brian L. Black, Edward M. Rubin, Axel Visel, and Len A. Pennacchio. ChIP-seq identification of weakly conserved heart enhancers. *Nature Genetics*, 42(9):806–812, 2010.

[232] Dalit May, Matthew J. Blow, Tommy Kaplan, David J. McCulley, Brian C. Jensen, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Veena Afzal, Paul C. Simpson, Edward M. Rubin, Brian L. Black, James Bristow, Len A. Pennacchio, and Axel Visel. Large-scale discovery of enhancers from human heart tissue. *Nature Genetics*, 2012.

[233] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–285, 2011.

[234] Michael Z. Ludwig, Casey Bergman, Nipam H. Patel, and Martin KreLtman. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769):564–567, 2000.

[235] David N. Arnosti, Scott Barolo, Michael Levine, and Stephen Small. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development*, 122(1):205–214, 1996.

[236] Sang Hyun Song, Ae Ri Kim, Tobias Ragoczy, M. A. Bender, Mark Groudine, and Ann Dean. Multiple functions of Ldb1 required for β-globin activation during erythroid differentiation. *Blood*, 116(13):2356–2364, 2010.

[237] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[238] Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, Alex Reynolds, Eric Haugen, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Richard Sandstrom, Jeff Vierstra, Rajinder Kaul, and John

Stamatoyannopoulos. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584(7820):244–251, 2020.

[239] Vania Parelho, Suzana Hadjur, Mikhail Spivakov, Marion Leleu, Stephan Sauer, Heather C. Gregson, Adam Jarmuz, Claudia Canzonetta, Zoe Webster, Tatyana Nesterova, Bradley S. Cobb, Kyoko Yokomori, Niall Dillon, Luis Aragon, Amanda G. Fisher, and Matthias Merkenschlager. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell*, 132(3):422–433, 2008.

[240] Federico Abascal, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Bronwen Aken, and Jennifer A. Akiyama. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[241] M. Ryan Corces, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, Clarice Groeneveld, Christopher K. Wong, Seung Woo Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nathan C. Sheffield, Ina Felau, Mauro A.A. Castro, Benjamin P. Berman, Louis M. Staudt, Jean C. Zenklusen, Peter W. Laird, Christina Curtis, William J. Greenleaf, and Howard Y. Chang. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), 2018.

[242] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[243] Gabriel E. Zentner, Paul J. Tesar, and Peter C. Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, 2011.

[244] Stefan Bonn, Robert P. Zinzen, Charles Girardot, E. Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyåski, Andrew Riddell, and Eileen E.M. Furlong. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156, 2012.

[245] Menno P. Creyghton, Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, Michael A. Lodato, Garrett M. Frampton, Phillip A. Sharp, Laurie A. Boyer, Richard A. Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 12 2010.

[246] Ali Shilatifard. The COMPASS family of histone H3K4 methylases: Mechanisms of regulation in development and disease pathogenesis. *Annual Review of Biochemistry*, 81:65–95, 2012.

[247] Christie C. Sze and Ali Shilatifard. MLL3/MLL4/COMPASS family on epigenetic regulation of enhancer function and cancer. *Cold Spring Harbor Perspectives in Medicine*, 6(11):1–16, 2016.

[248] Thomas A. Milne, Scott D. Briggs, Hugh W. Brock, Mary Ellen Martin, Denise Gibbs, C. David Allis, and Jay L. Hess. MLL targets SET domain methyltransferase activity to Hox gene promoters. *Molecular Cell*, 10(5):1107–1117, 2002.

[249] Tatsuya Nakamura, Toshiki Mori, Shinichiro Tada, Wladyslaw Krajewski, Tanya Rozovskaia, Richard Wassell, Garrett Dubois, Alexander Mazo, Carlo M. Croce, and Eli Canaani. ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Molecular Cell*, 10(5):1119–1128, 2002.

[250] Deqing Hu, Xin Gao, Marc A. Morgan, Hans-Martin Herz, Edwin R. Smith, and Ali Shilatifard. The MLL3/MLL4 Branches of the COMPASS Family Function as Major Histone H3K4 Monomethylases at Enhancers. *Molecular and Cellular Biology*, 33(23):4745–4754, 2013.

[251] Hans Martin Herz, Man Mohan, Alexander S. Garruss, Kaiwei Liang, Yoh hei Takahashi, Kristen Mickey, Olaf Voets, C. Peter Verrijzer, and Ali Shilatifard. Enhancer-associated H3K4 monomethylation by trithorax-related, the drosophila homolog of mammalian MLL3/MLL4. *Genes and Development*, 26(23):2604–2620, 2012.

[252] Ji Eun Lee, Chaochen Wang, Shiliyang Xu, Young Wook Cho, Lifeng Wang, Xuesong Feng, Anne Baldridge, Vittorio Sartorelli, Lenan Zhuang, Weiqun Peng,

and Kai Ge. H3K4 mono- And di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife*, 2013(2):1–25, 2013.

[253] Minna U. Kaikkonen, Nathanael J. Spann, Sven Heinz, Casey E. Romanoski, Karmel A. Allison, Joshua D. Stender, Hyun B. Chun, David F. Tough, Rab K. Prinjha, Christopher Benner, and Christopher K. Glass. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell*, 51(3):310–325, 2013.

[254] Ana Ortega-Molina, Isaac W. Boss, Andres Canela, Heng Pan, Yanwen Jiang, Chunying Zhao, Man Jiang, Deqing Hu, Xabier Agirre, Itamar Niesvizky, Ji Eun Lee, Hua Tang Chen, Daisuke Ennishi, David W. Scott, Anja Mottok, Christoffer Hother, Shichong Liu, Xing Jun Cao, Wayne Tam, Rita Shaknovich, Benjamin A. Garcia, Randy D. Gascoyne, Kai Ge, Ali Shilatifard, Olivier Elemento, Andre Nussenzweig, Ari M. Melnick, and Hans Guido Wendel. The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nature Medicine*, 21(10):1199–1208, 2015.

[255] Siang Yun Ang, Alec Uebersohn, C. Ian Spencer, Yu Huang, Ji Eun Lee, Kai Ge, and Benoit G. Bruneau. KMT2D regulates specific programs in heart development via histone H3 lysine 4 di-methylation. *Development* (*Cambridge*), 143(5):810–821, 2016.

[256] Kyoung Jae Won, Zheng Xu, Xian Zhang, John W. Whitaker, Robert Shoemaker, Bing Ren, Yang Xu, and Wei Wang. Global identification of transcriptional regulators of pluripotency and differentiation in embryonic stem cells. *Nucleic Acids Research*, 40(17):8199–8209, 2012.

[257] Beverley M. Dancy and Philip A. Cole. Protein lysine acetylation by p300/CBP. *Chemical Reviews*, 115(6):2419–2452, 2015.

[258] Jareth C. Wolfe, Liudmila A. Mikheeva, Hani Hagras, and Nicolae Radu Zabet. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. *Genome Biology*, 22(1):1–23, 2021.

[259] Gillian C.A. Taylor, Ragnhild Eskeland, Betül Hekimoglu-Balkan, Madapura M. Pradeepa, and Wendy A. Bickmore. H4K16 acetylation marks active genes and en-

hancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Research*, 23(12):2053–2065, 2013.

[260] Madapura M. Pradeepa, Graeme R. Grimes, Yatendra Kumar, Gabrielle Olley, Gillian C.A. Taylor, Robert Schneider, and Wendy A. Bickmore. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics*, 48(6):681–686, 2016.

[261] Amanuel Tafessu and Laura A. Banaszynski. Establishment and function of chromatin modification at enhancers: Chromatin Landscape at Enhancers. *Open Biology*, 10(10), 2020.

[262] Tiantian Zhang, Zhuqiang Zhang, Qiang Dong, Jun Xiong, and Bing Zhu. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21(1):1–7, 2020.

[263] Aditya Sankar, Faizaan Mohammad, Arun Kumar Sundaramurthy, Hua Wang, Mads Lerdrup, Tulin Tatar, and Kristian Helin. Histone editing elucidates the functional roles of H3K27 methylation and acetylation in mammals. *Nature Genetics*, 54(6):754–760, 2022.

[264] Sara Cruz-Molina, Patricia Respuela, Christina Tebartz, Petros Kolovos, Milos Nikolic, Raquel Fueyo, Wilfred F.J. van Ijcken, Frank Grosveld, Peter Frommolt, Hisham Bazzi, and Alvaro Rada-Iglesias. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell*, 20(5):689–705, 2017.

[265] Tomas Pachano, Víctor Sánchez-Gaya, Thais Ealo, Maria Mariner-Faulí, Tore Bleckwehl, Helena G. Asenjo, Patricia Respuela, Sara Cruz-Molina, María Muñoz-San Martín, Endika Haro, Wilfred F.J. van IJcken, David Landeira, and Alvaro Rada-Iglesias. Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nature Genetics*, 53(7):1036–1049, 2021.

[266] Na Li, Yuanyuan Li, Jie Lv, Xiangdong Zheng, Hong Wen, Hongjie Shen, Guangjing Zhu, Tsai Yu Chen, Shilpa S. Dhar, Pu Yeh Kan, Zhibin Wang, Ramin Shiekhattar, Xiaobing Shi, Fei Lan, Kaifu Chen, Wei Li, Haitao Li, and Min Gyu Lee. ZMYND8 Reads the Dual Histone Mark H3K4me1-H3K14ac to Antagonize the Expression of Metastasis-Linked Genes. *Molecular Cell*, 63(3):470–484, 2016.

[267] Verónica Delgado-Benito, Daniel B. Rosen, Qiao Wang, Anna Gazumyan, Joy A. Pai, Thiago Y. Oliveira, Devakumar Sundaravinayagam, Wenzhu Zhang, Matteo Andreani, Lisa Keller, Kyong Rim Kieffer-Kwon, Aleksandra Pękowska, Seolkyoung Jung, Madlen Driesner, Roman I. Subbotin, Rafael Casellas, Brian T. Chait, Michel C. Nussenzweig, and Michela Di Virgilio. The Chromatin Reader ZMYND8 Regulates Igh Enhancers to Promote Immunoglobulin Class Switch Recombination. *Molecular Cell*, 72(4):636–649, 2018.

[268] Pavel Savitsky, Tobias Krojer, Takao Fujisawa, Jean Philippe Lambert, Sarah Picaud, Chen Yi Wang, Erin K. Shanle, Krzysztof Krajewski, Hans Friedrichsen, Alexander Kanapin, Colin Goding, Matthieu Schapira, Anastasia Samsonova, Brian D. Strahl, Anne Claude Gingras, and Panagis Filippakopoulos. Multivalent Histone and DNA Engagement by a PHD/BRD/PWWP Triple Reader Cassette Recruits ZMYND8 to K14ac-Rich Chromatin. *Cell Reports*, 17(10):2724–2737, 2016.

[269] Andrea Local, Hui Huang, Claudio P. Albuquerque, Namit Singh, Ah Young Lee, Wei Wang, Chaochen Wang, Judy E. Hsia, Andrew K. Shiau, Kai Ge, Kevin D. Corbett, Dong Wang, Huilin Zhou, and Bing Ren. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature Genetics*, 50(1):73–82, 2018.

[270] Mikhail Spivakov and Amanda G. Fisher. Epigenetic signatures of stem-cell identity. *Nature Reviews Genetics*, 8(4):263–271, 2007.

[271] Roland Jäger, Gabriele Migliorini, Marc Henrion, Radhika Kandaswamy, Helen E. Speedy, Andreas Heindl, Nicola Whiffin, Maria J. Carnicer, Laura Broome, Nicola Dryden, Takashi Nagano, Stefan Schoenfelder, Martin Enge, Yinyin Yuan, Jussi Taipale, Peter Fraser, Olivia Fletcher, and Richard S. Houlston. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, 6:1–9, 2015.

[272] Martin Franke, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, 2016.

[273] Guillaume Andrey, Robert Schöpflin, Ivana Jerković, Verena Heinrich, Daniel M. Ibrahim, Christina Paliou, Myriam Hochradel, Bernd Timmermann, Stefan Haas, Martin Vingron, and Stefan Mundlos. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Research*, 27(2):223–233, 2017.

[274] Stefan Schoenfelder, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola Maria Javierre, Takashi Nagano, Yulia Katsman, Moorthy Sakthidevi, Steven W. Wingett, Emilia Dimitrova, Andrew Dimond, Lucas B. Edelman, Sarah Elderkin, Kristina Tabbada, Elodie Darbo, Simon Andrews, Bram Herman, Andy Higgs, Emily LeProust, Cameron S. Osborne, Jennifer A. Mitchell, Nicholas M. Luscombe, and Peter Fraser. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4):582–597, 2015.

[275] Borbala Mifsud, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, Simon Andrews, William Grey, Philip A. Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily Leproust, George A. Follows, Peter Fraser, Nicholas M. Luscombe, and Cameron S. Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, 2015.

[276] Pelin Sahlén, Ilgar Abdullayev, Daniel Ramsköld, Liudmila Matskova, Nemanja Rilakovic, Britta Lötstedt, Thomas J. Albert, Joakim Lundeberg, and Rickard Sandberg. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, 16(1):1–13, 2015.

[277] Paula Freire-Pritchett, Stefan Schoenfelder, Csilla Várnai, Steven W. Wingett, Jonathan Cairns, Amanda J. Collier, Raquel García-Vílchez, Mayra Furlan-Magaril, Cameron S. Osborne, Peter Fraser, Peter J. Rugg-Gunn, and Mikhail Spivakov. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife*, 6:1–26, 2017.

[278] Peter Chovanec, Amanda J. Collier, Christel Krueger, Csilla Várnai, Claudia I. Semprich, Stefan Schoenfelder, Anne E. Corcoran, and Peter J. Rugg-Gunn. Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. *Nature Communications*, 12(1):1–18, 2021.

[279] Rasmus Siersbæk, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen, Emilie Kristine Bagge, Jonathan Cairns, Steven William Wingett, Sofie Traynor, Mikhail Spivakov, Peter Fraser, and Susanne Mandrup. Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Molecular Cell*, 66(3):420–435, 2017.

[280] Adam J. Rubin, Brook C. Barajas, Mayra Furlan-Magaril, Vanessa Lopez-Pajares, Maxwell R. Mumbach, Imani Howard, Daniel S. Kim, Lisa D. Boxer, Jonathan Cairns, Mikhail Spivakov, Steven W. Wingett, Minyi Shi, Zhixin Zhao, William J. Greenleaf, Anshul Kundaje, Michael Snyder, Howard Y. Chang, Peter Fraser, and Paul A. Khavari. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nature Genetics*, 49(10):1522–1528, 2017.

[281] Biola M. Javierre, Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, Mikhail Spivakov, Peter Fraser, Oliver S. Burren, Antony J. Cutler, John A. Todd, Chris Wallace, Steven P. Wilder, Roman Kreuzhuber, Myrto Kostadima, Daniel R. Zerbino, Oliver Stegle, Frances Burden, Samantha Farrow, Karola Rehnström, Kate Downes, Luigi Grassi, Willem H. Ouwehand, Mattia Frontini, Steven M. Hill, Fan Wang, Hendrik G. Stunnenberg, Joost H. Martens, Bowon Kim, Nilofar Sharifi, Eva M. Janssen-Megens, Marie Laure Yaspo, Matthias Linser, Alexander Kovacsovics, Laura Clarke, David Richardson, Avik Datta, and Paul Flicek. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384, 2016.

[282] Oliver S. Burren, Arcadio Rubio García, Biola Maria Javierre, Daniel B. Rainbow, Jonathan Cairns, Nicholas J. Cooper, John J. Lambourne, Ellen Schofield, Xaquin Castro Dopico, Ricardo C. Ferreira, Richard Coulson, Frances Burden, Sophia P. Rowlston, Kate Downes, Steven W. Wingett, Mattia Frontini, Willem H. Ouwehand, Peter Fraser, Mikhail Spivakov, John A. Todd, Linda S. Wicker, Antony J. Cutler, and Chris Wallace. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biology*, 18(1):1–19, 2017.

[283] Romina Petersen, John J. Lambourne, Biola M. Javierre, Luigi Grassi, Roman Kreuzhuber, Dace Ruklisa, Isabel M. Rosa, Ana R. Tomé, Heather Elding, Johanna P.

Van Geffen, Tao Jiang, Samantha Farrow, Jonathan Cairns, Abeer M. Al-Subaie, Sofie Ashford, Antony Attwood, Joana Batista, Heleen Bouman, Frances Burden, Fizzah A. Choudry, Laura Clarke, Paul Flicek, Stephen F. Garner, Matthias Haimel, Carly Kempster, Vasileios Ladopoulos, An Sofie Lenaerts, Paulina M. Materek, Harriet McKinney, Stuart Meacham, Daniel Mead, Magdolna Nagy, Christopher J. Penkett, Augusto Rendon, Denis Seyres, Benjamin Sun, Salih Tuna, Marie Elise Van Der Weide, Steven W. Wingett, Joost H. Martens, Oliver Stegle, Sylvia Richardson, Ludovic Vallier, David J. Roberts, Kathleen Freson, Lorenz Wernisch, Hendrik G. Stunnenberg, John Danesh, Peter Fraser, Nicole Soranzo, Adam S. Butterworth, Johan W. Heemskerk, Ernest Turro, Mikhail Spivakov, Willem H. Ouwehand, William J. Astle, Kate Downes, Myrto Kostadima, and Mattia Frontini. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nature Communications*, 8(May), 2017.

[284] Kevin Litchfield, Max Levy, Giulia Orlando, Chey Loveday, Philip J. Law, Gabriele Migliorini, Amy Holroyd, Peter Broderick, Robert Karlsson, Trine B. Haugen, Wenche Kristiansen, Jérémie Nsengimana, Kerry Fenwick, Ioannis Assiotis, Zsofia Kote-Jarai, Alison M. Dunning, Kenneth Muir, Julian Peto, Rosalind Eeles, Douglas F. Easton, Darshna Dudakia, Nick Orr, Nora Pashayan, D. Timothy Bishop, Alison Reid, Robert A. Huddart, Janet Shipley, Tom Grotmol, Fredrik Wiklund, Richard S. Houlston, and Clare Turnbull. Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor. *Nature Genetics*, 49(7):1133–1140, 2017.

[285] Lindsey E. Montefiori, Debora R. Sobreira, Noboru J. Sakabe, Ivy Aneas, Amelia C. Joslin, Grace T. Hansen, Grazyna Bozek, Ivan P. Moskowitz, Elizabeth M. McNally, and Marcelo A. Nóbrega. A promoter interaction map for cardiovascular disease genetics. *eLife*, 7:1–35, 2018.

[286] Mun Kit Choy, Biola M. Javierre, Simon G. Williams, Stephanie L. Baross, Yingjuan Liu, Steven W. Wingett, Artur Akbarov, Chris Wallace, Paula Freire-Pritchett, Peter J. Rugg-Gunn, Mikhail Spivakov, Peter Fraser, and Bernard D. Keavney. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nature Communications*, 9(1), 2018.

[287] Helen Ray-Jones, Kate Duffus, Amanda McGovern, Paul Martin, Chenfu Shi, Jenny

Hankinson, Oliver Gough, Annie Yarwood, Andrew P. Morris, Antony Adamson, Christopher Taylor, James Ding, Vasanthi Priyadarshini Gaddi, Yao Fu, Patrick Gaffney, Gisela Orozco, Richard B. Warren, and Steve Eyre. Chromatin-based techniques map DNA interaction landscapes in psoriasis susceptibility loci and highlight KLF4 as a target gene in 9q31. *bioRxiv*, pages 1–20, 2019.

[288] Paul Martin, James Ding, Kate Duffus, Vasanthi Priyadarshini Gaddi, Amanda Mcgovern, Helen Ray-Jones, Annie Yarwood, Jane Worthington, Anne Barton, and Gisela Orozco. Chromatin interactions reveal novel gene targets for drug repositioning in rheumatic diseases. *Annals of the Rheumatic Diseases*, 78(8):1127–1134, 2019.

[289] Peng Hua, Mohsin Badat, Lars L.P. Hanssen, Lance D. Hentges, Nicholas Crump, Damien J. Downes, Danuta M. Jeziorska, A. Marieke Oudelaar, Ron Schwessinger, Stephen Taylor, Thomas A. Milne, Jim R. Hughes, Doug R. Higgs, and James O.J. Davies. *Defining genome architecture at base-pair resolution*, volume 595. Springer US, 2021.

[290] Jorg Bungert, Utpal Dave, Kim Chew Lim, Ken H. Lieuw, Jordan A. Shavit, Qinghui Liu, and James Douglas Engel. Synergistic regulation of human β-globin gene switching by locus control region elements HS3 and HS4. *Genes and Development*, 9(24):3083–3096, 1995.

[291] Mark Wijgerde, Frank Grosveld, and Peter Fraser. Transcription complex stability and chromatin dynamics in viv. *Nature*, 377(6546):209–213, 1995.

[292] Michael Bulger and Mark Groudine. Looping versus linking: Toward a model for long-distance gene activation. *Genes and Development*, 13(19):2465–2477, 1999.

[293] Dale Dorsett. Distant liaisons: Long-range enhancer-promoter interactions in Drosophila. *Current Opinion in Genetics and Development*, 9(5):505–514, 1999.

[294] James Douglas Engel and Keiji Tanimoto. Looping, linking, and chromatin activity: New insights into β-globin locus regulation. *Cell*, 100(5):499–502, 2000.

[295] Mark Ptashne. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322(6081):697–701, 1986.

[296] Adam G. West, Miklos Gaszner, and Gary Felsenfeld. Insulators: Many functions, many mechanisms. *Genes and Development*, 16(3):271–288, 2002.

[297] Niall Dillon and Pierangela Sabbattini. Functional gene expression domains: Defining the functional unit of eukaryotic gene regulation. *BioEssays*, 22(7):657–665, 2000.

[298] David Carter, Lyubomira Chakalova, Cameron S. Osborne, Yan feng Dai, and Peter Fraser. Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, 32(4):623–626, 2002.

[299] Bas Tolhuis, Robert Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter De Laat. Looping and interaction between hypersensitive sites in the active β-globin locus. *Molecular Cell*, 10(6):1453–1465, 2002.

[300] Sergio Martin Espinola, Markus Götz, Maelle Bellec, Olivier Messina, Jean Bernard Fiche, Christophe Houbron, Matthieu Dejean, Ingolf Reim, Andrés M. Cardozo Gizzi, Mounia Lagha, and Marcelo Nollmann. Cis-regulatory chromatin loops arise before TADs and gene activation, and are independent of cell fate during early Drosophila development. *Nature Genetics*, 53(4):477–486, 2021.

[301] Tessa M Popay and Jesse R Dixon. Coming full circle: on the origin and evolution of the looping model for enhancer- promoter communication. *Journal of Biological Chemistry*, 2022.

[302] Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339, 2011.

[303] Anil Panigrahi and Bert W. O'Malley. Mechanisms of enhancer action: the known and the unknown. *Genome Biology*, 22(1):1–30, 2021.

[304] Robert A. Beagrie and Ana Pombo. Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. *BioEssays*, 38(9):881–893, 2016.

[305] Mark Ptashne and Alexander Gann. Transcriptional activation by recruitment. *Nature*, 386:569–577, 1997.

[306] Bryan Lemon and Robert Tjian. Orchestrated response: A symphony of transcription factors for gene control. *Genes and Development*, 14(20):2551–2569, 2000.

[307] D Herschlag and F B Johnson. Synergism in transcriptional activation: a kinetic view. *Genes and Deve*, 7:173–179, 1993.

[308] Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, 24(3):390–400, 2014.

[309] Fei Sun, Constantinos Chronis, Michael Kronenberg, Xiao Fen Chen, Trent Su, Fides D. Lay, Kathrin Plath, Siavash K. Kurdistani, and Michael F. Carey. Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. *Molecular Cell*, 73(2):250–263, 2019.

[310] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[311] Uirá Souto Melo, Robert Schöpflin, Rocio Acuna-Hidalgo, Martin Atta Mensah, Björn Fischer-Zirnsak, Manuel Holtgrewe, Marius Konstantin Klever, Seval Türkmen, Verena Heinrich, Ilina Datkhaeva Pluym, Eunice Matoso, Sérgio Bernardo de Sousa, Pedro Louro, Wiebke Hülsemann, Monika Cohen, Andreas Dufke, Anna Latos-Bieleńska, Martin Vingron, Vera Kalscheuer, Fabiola Quintero-Rivera, Malte Spielmann, and Stefan Mundlos. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *American Journal of Human Genetics*, 106(6):872–884, 2020.

[312] Magdalena Laugsch, Michaela Bartusel, Rizwan Rehimi, Hafiza Alirzayeva, Agathi Karaolidou, Giuliano Crispatzu, Peter Zentis, Milos Nikolic, Tore Bleckwehl, Petros Kolovos, Wilfred F.J. van Ijcken, Tomo Šarić, Katrin Koehler, Peter Frommolt, Katherine Lachlan, Julia Baptista, and Alvaro Rada-Iglesias. Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs. *Cell Stem Cell*, 24(5):736–752, 2019.

[313] Alexandra Despang, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing Lee Chan, Bernd Timmermann, Lars Wittler, Martin Vingron, Stefan Mundlos, and Daniel M. Ibrahim. Functional dissection of the

Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics*, 51(8):1263–1271, 2019.

[314] Yad Ghavi-Helm, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E.M. Furlong. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, 51(8):1272–1282, 2019.

[315] Michiel J. Thiecke, Gordana Wutz, Matthias Muhar, Wen Tang, Stephen Bevan, Valeriya Malysheva, Roman Stocsits, Tobias Neumann, Johannes Zuber, Peter Fraser, Stefan Schoenfelder, Jan Michael Peters, and Mikhail Spivakov. Cohesin-Dependent and -Independent Mechanisms Mediate Chromosomal Contacts between Promoters and Enhancers. *Cell Reports*, 32(3), 2020.

[316] Kyle P. Eagen, Erez Lieberman Aiden, and Roger D. Kornberg. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences of the United States of America*, 114(33):8764–8769, 2017.

[317] Yuki Ogiyama, Bernd Schuettengruber, Giorgio L. Papadopoulos, Jia Ming Chang, and Giacomo Cavalli. Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development. *Molecular Cell*, 71(1):73–88, 2018.

[318] Abraham S. Weintraub, Charles H. Li, Alicia V. Zamudio, Alla A. Sigova, Nancy M. Hannett, Daniel S. Day, Brian J. Abraham, Malkiel A. Cohen, Behnam Nabet, Dennis L. Buckley, Yang Eric Guo, Denes Hnisz, Rudolf Jaenisch, James E. Bradner, Nathanael S. Gray, and Richard A. Young. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, 171(7):1573–1588, 2017.

[319] Swneke D. Bailey, Xiaoyang Zhang, Kinjal Desai, Malika Aid, Olivia Corradin, Richard Cowper-Sallari, Batool Akhtar-Zaidi, Peter C. Scacheri, Benjamin Haibe-Kains, and Mathieu Lupien. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, 2, 2015.

[320] B. Y. Ye, W. L. Shen, D. Wang, P. Li, Z. Zhang, M. L. Shi, Y. Zhang, F. X. Zhang, and Z. H. Zhao. ZNF143 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. *Molecular Biology*, 50(3):431–437, 2016.

[321] Jonathan A. Beagan, Michael T. Duong, Katelyn R. Titus, Linda Zhou, Zhendong Cao, Jingjing Ma, Caroline V. Lachanski, Daniel R. Gillis, and Jennifer E. Phillips-Cremins. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Research*, 27(7):1139–1152, 2017.

[322] Nastaran Heidari, Douglas H. Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q. Zhang, and Michael P. Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Research*, 24(12):1905–1917, 2014.

[323] Linda W. Jurata, Samuel L. Pfaff, and Gordon N. Gill. The nuclear LIM domain interactor NLI mediates homo- and heterodimerization of LIM domain transcription factors. *Journal of Biological Chemistry*, 273(6):3152–3157, 1998.

[324] Sang Hyun Song, Chunhui Hou, and Ann Dean. A Positive Role for NLI/Ldb1 in Long-Range β-Globin Locus Control Region Function. *Molecular Cell*, 28(5):810–822, 2007.

[325] Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244, 2012.

[326] Alessandro Magli, June Baik, Pruthvi Pota, Carolina Ortiz Cordero, Il Youp Kwak, Daniel J. Garry, Paul E. Love, Brian D. Dynlacht, and Rita C.R. Perlingeiro. Pax3 cooperates with Ldb1 to direct local chromosome architecture during myogenic lineage specification. *Nature Communications*, 10(1), 2019.

[327] Kevin Monahan, Adan Horta, and Stavros Lomvardas. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature*, 565(7740):448–453, 2019.

[328] Wulan Deng, Jeremy W. Rupon, Ivan Krivega, Laura Breda, Irene Motta, Kristen S. Jahn, Andreas Reik, Philip D. Gregory, Stefano Rivella, Ann Dean, and Gerd A. Blobel. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell*, 158(4):849–860, 2014.

[329] Ivan Krivega, Ryan K. Dale, and Ann Dean. Role of LDB1 in the transition from

chromatin looping to transcription activation. *Genes and Development*, 28(12):1278–1290, 2014.

[330] Bas van Steensel and Eileen E.M. Furlong. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology*, 20(6):327–337, 2019.

[331] Hongtao Chen, Michal Levo, Lev Barinov, Miki Fujioka, James B. Jaynes, and Thomas Gregor. Dynamic interplay between enhancer–promoter topology and gene activity. *Nature Genetics*, 50(9):1296–1303, 2018.

[332] Nezha S. Benabdallah, Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular Cell*, 76(3):473–484, 2019.

[333] Jeffrey M. Alexander, Juan Guan, Bingkun Li, Lenka Maliskova, Michael Song, Yin Shen, Bo Huang, Stavros Lomvardas, and Orion D. Weiner. Live-cell imaging reveals enhancer-dependent sox2 transcription in the absence of enhancer proximity. *eLife*, 2019.

[334] Elizabeth Ing-Simmons, Roshan Vaid, Xin Yang Bing, Michael Levine, Mattias Mannervik, and Juan M. Vaquerizas. Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning. *Nature Genetics*, 53(4):487–499, 2021.

[335] Anthony A. Hyman, Christoph A. Weber, and Frank Jülicher. Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology*, 30:39–58, 2014.

[336] Salman F. Banani, Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5):285–298, 2017.

[337] Pilong Li, Sudeep Banjade, Hui Chun Cheng, Soyeon Kim, Baoyu Chen, Liang Guo, Marc Llaguno, Javoris V. Hollingsworth, David S. King, Salman F. Banani, Paul S. Russo, Qiu Xing Jiang, B. Tracy Nixon, and Michael K. Rosen. Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340, 2012.

[338] Bryan A. Gibson, Lynda K. Doolittle, Maximillian W.G. Schneider, Liv E. Jensen, Nathan Gamarra, Lisa Henry, Daniel W. Gerlich, Sy Redding, and Michael K. Rosen. Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell*, 179(2):470–484, 2019.

[339] S. Sanulli, M. J. Trnka, V. Dharmarajan, R. W. Tibble, B. D. Pascal, A. L. Burlingame, P. R. Griffin, J. D. Gross, and G. J. Narlikar. HP1 reshapes nucleosome core to promote phase separation of heterochromatin. *Nature*, 575(7782):390–394, 2019.

[340] Yi Zhang and Tatiana G. Kutateladze. Liquid–liquid phase separation is an intrinsic physicochemical property of chromatin. *Nature Structural and Molecular Biology*, 26(12):1085–1086, 2019.

[341] Mustafa Mir, Michael R. Stadler, Stephan A. Ortiz, Colleen E. Hannon, Melissa M. Harrison, Xavier Darzacq, and Michael B. Eisen. Dynamic multifactor hubs interact transiently with sites of active transcription in drosophila embryos. *eLife*, 7(Dv):1–27, 2018.

[342] Jacob H. Hanna, Krishanu Saha, and Rudolf Jaenisch. Pluripotency and cellular reprogramming: Facts, hypotheses, unresolved issues. *Cell*, 143(4):508–525, 2010.

[343] Jennifer Nichols and Austin Smith. Pluripotency in the embryo and in culture. *Cold Spring Harbor Perspectives in Biology*, 4(8), 2012.

[344] Jamie A. Hackett and M. Azim Surani. Regulatory principles of pluripotency: From the ground state up. *Cell Stem Cell*, 15(4):416–430, 2014.

[345] Leehee Weinberger, Muneef Ayyash, Noa Novershtern, and Jacob H. Hanna. Dynamic stem cell states: Naive to primed pluripotency in rodents and humans. *Nature Reviews Molecular Cell Biology*, 17(3):155–169, 2016.

[346] Peter Braude, Virginia Bolton, and Stephen Moore. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature*, 332:459–461, 1988.

[347] Paul Blakeley, Norah M E Fogarty, Ignacio Valle, Sissy E Wamaitha, Tim Xiaoming Hu, Kay Elder, Philip Snell, Leila Christie, Paul Robson, Kathy K Niakan, Paul Blakeley, Norah M E Fogarty, Ignacio Valle, Sissy E Wamaitha, Tim Xiaoming Hu,

Kay Elder, Philip Snell, Leila Christie, Paul Robson, and Kathy K Niakan. Erratum to Defining the three cell lineages of the human blastocyst by single-cell RNA-seq (Development, (2015) 142, 3151-3165). *Development (Cambridge)*, 142(20):3613, 2015.

[348] R.G. Edwards, J.M. Purdyab, P.C. Steptoe, and D.E. Walters. The growth of human preimplantation embryos in vitro. *American Journal of Obstetrics and Gynecology*, 141(4):408–416, 1981.

[349] George Nikas, Asangla Ao, Robert M.L. Winston, and Alan H. Handyside. Compaction and surface polarity in the human embryo in vitro. *Biology of Reproduction*, 55(1):32–37, 1996.

[350] P. C. STEPTOE, R. G. EDWARDS, and J. M. PURDY. Human Blastocysts grown in Culture. *Nature*, 229:132–133, 1971.

[351] Richard Lavenham Gardner. Clonal analysis of early mammalian development. *Royal Society*, 312(1153), 1985.

[352] Richard Lavenham Gardner and J Rossant. Investigation of the fate of 4-5 day post-coitum mouse inner cell mass cells by blastocyst injection. *J Embryol Exp Morphol*, 52:141–152, 1979.

[353] R L Gardner, V E Papaioannou, and S C Barton. Origin of the ectoplacental cone and secondary giant cells in mouse blastocysts reconstituted from isolated trophoblast and inner cell mass. *J Embryol Exp Morphol*, 30(3):561–572, 1973.

[354] Birgit Gellersen and Jan J. Brosens. Cyclic decidualization of the human endometrium in reproductive health and failure. *Endocrine Reviews*, 35(6):851–905, 2014.

[355] Sissy E. Wamaitha and Kathy K. Niakan. *Human Pre-gastrulation Development*, volume 128. Elsevier Inc., 2018.

[356] A T HERTIG, J ROCK, and E C ADAMS. A description of 34 human ova within the first 17 days of development. *Am J Anat*, 98(3):435–493, 1956.

[357] R O'Rahilly and F. Muller. Developmental stages in human embryos. *Carnegie Institution of Washington*, 1987.

[358] Kathy K. Niakan, Jinnuo Han, Roger A. Pedersen, Carlos Simon, and Renee A.Reijo Pera. Human pre-implantation embryo development. *Development*, 139(5):829–841, 2012.

[359] Behrouz Aflatoonian, Ludmila Ruban, Shamsul Shamsuddin, Duncan Baker, Peter Andrews, and Harry Moore. Generation of Sheffield (Shef) human embryonic stem cell lines using a microdrop culture system. *In Vitro Cellular and Developmental Biology - Animal*, 46(3-4):236–241, 2010.

[360] Chad A. Cowan, Irina Klimanskaya, Jill McMahon, Jocelyn Atienza, Jeannine Witmyer, Jacob P. Zucker, Shunping Wang, Cynthia C. Morton, Andrew P. McMahon, Doug Powers, and Douglas A. Melton. Derivation of Embryonic Stem-Cell Lines from Human Blastocysts. *New England Journal of Medicine*, 350(13):1353–1356, 2004.

[361] Maisam Mitalipova, John Calhoun, Soojung Shin, David Wininger, Thomas Schulz, Scott Noggle, Alison Venable, Ian Lyons, Allan Robins, and Steven Stice. Human Embryonic Stem Cell Lines Derived from Discarded Embryos. *Stem Cells*, 21(5):521–526, 2003.

[362] Hirofumi Suemori, Kentaro Yasuchika, Kouichi Hasegawa, Tsuyoshi Fujioka, Norihiro Tsuneyoshi, and Norio Nakatsuji. Efficient establishment of human embryonic stem cell lines and long-term maintenance with stable karyotype by enzymatic bulk passage. *Biochemical and Biophysical Research Communications*, 345(3):926–932, 2006.

[363] Evans M J and Kaufman M H. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(July):154–156, 1981.

[364] G. R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 78(12 II):7634–7638, 1981.

[365] Allan Bradley, Martin Evans, Matthew H. Kaufman, and Elizabeth Robertson. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*, 309(5965):255–256, 1984.

[366] James A. Thomson, J Itskovitz-Eldor, S S Shapiro, M A Waknitz, J J Swiergiel, V S Marshall, and J M Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282((5391)):1145–7, 1998.

[367] Ludovic Vallier, Morgan Alexander, and Roger A. Pedersen. Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *Journal of Cell Science*, 118(19):4495–4509, 2005.

[368] Gillian M. Beattie, Ana D. Lopez, Nathan Bucay, Andrew Hinton, Meri T. Firpo, Charles C. King, and Alberto Hayek. Activin A Maintains Pluripotency of Human Embryonic Stem Cells in the Absence of Feeder Layers. *Stem Cells*, 23(4):489–495, 2005.

[369] Lisa M. Hoffman, Lisa Hall, Jennifer L. Batten, Holly Young, Dheerja Pardasani, E. Edward Baetge, Jeanne Lawrence, and Melissa K. Carpenter. X-Inactivation Status Varies in Human Embryonic Stem Cell Lines. *Stem Cells*, 23(10):1468–1478, 2005.

[370] Paul J. Tesar, Josh G. Chenoweth, Frances A. Brook, Timothy J. Davies, Edward P. Evans, David L. Mack, Richard L. Gardner, and Ronald D.G. McKay. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, 2007.

[371] I. Gabrielle M. Brons, Lucy E. Smithers, Matthew W.B. Trotter, Peter Rugg-Gunn, Bowen Sun, Susana M. Chuva De Sousa Lopes, Sarah K. Howlett, Amanda Clarkson, Lars Ahrlund-Richter, Roger A. Pedersen, and Ludovic Vallier. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448(7150):191–195, 2007.

[372] Jennifer Nichols and Austin Smith. Naive and Primed Pluripotent States. *Cell Stem Cell*, 4(6):487–492, 2009.

[373] Ge Guo, Ferdinand von Meyenn, Maria Rostovskaya, James Clarke, Sabine Dietmann, Duncan Baker, Anna Sahakyan, Samuel Myers, Paul Bertone, Wolf Reik, Kathrin Plath, and Austin Smith. Epigenetic resetting of human pluripotency. *Development (Cambridge, England)*, 145(8), 2018.

[374] Gue Guo, Ferdinand von Meyenn, Fatima Santos, Yaoyao Chen, Wolf Reik, Paul Bertone, Austin Smith, and Jennifer Nichols. Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports*, 6(4):437–446, 2016.

[375] Tomonori Nakamura, Ikuhiro Okamoto, Kotaro Sasaki, Yukihiro Yabuta, Chizuru Iwatani, Hideaki Tsuchiya, Yasunari Seita, Shinichiro Nakamura, Takuya Yamamoto, and Mitinori Saitou. A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature*, 537(7618):57–62, 2016.

[376] Giuliano G. Stirparo, Thorsten Boroviak, Ge Guo, Jennifer Nichols, Austin Smith, and Paul Bertone. Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. *Development (Cambridge, England)*, 145(15), 2018.

[377] Yasuhiro Takashima, G. Guo, Remco Loos, Jennifer Nichols, Gabriella Ficz, Felix Krueger, David Oxley, Fatima Santos, James Clarke, William Mansfield, Wolf Reik, Paul Bertone, and Austin Smith. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell*, 158(6):1254–1269, 2014.

[378] Amanda J. Collier and Peter J. Rugg-Gunn. Identifying Human Naïve Pluripotent Stem Cells Evaluating State-Specific Reporter Lines and Cell-Surface Markers. *BioEssays*, 40(5):1–12, 2018.

[379] Thorold W. Theunissen, Marc Friedli, Yupeng He, Evarist Planet, Ryan C. O'Neil, Styliani Markoulaki, Julien Pontis, Haoyi Wang, Alexandra Iouranova, Michaël Imbeault, Julien Duc, Malkiel A. Cohen, Katherine J. Wert, Rosa Castanon, Zhuzhu Zhang, Yanmei Huang, Joseph R. Nery, Jesse Drotar, Tenzin Lungjangwa, Didier Trono, Joseph R. Ecker, and Rudolf Jaenisch. Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell*, 19(4):502–515, 2016.

[380] William A. Pastor, Di Chen, Wanlu Liu, Rachel Kim, Anna Sahakyan, Anastasia Lukianchikov, Kathrin Plath, Steven E. Jacobsen, and Amander T. Clark. Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. *Cell Stem Cell*, 18(3):323–329, 2016.

[381] Mario Iurlaro, Ferdinand von Meyenn, and Wolf Reik. DNA methylation homeostasis in human and mouse development. *Current Opinion in Genetics and Development*, 43:101–109, 2017.

[382] Carol B. Ware. Concise Review: Lessons from Naıve Human Pluripotent Cells. *Stem Cells*, 35:35–41, 2016.

[383] Banushree Kumar and Simon J. Elsässer. Quantitative Multiplexed ChIP Reveals Global Alterations that Shape Promoter Bivalency in Ground State Embryonic Stem Cells. *Cell Reports*, 28(12):3274–3284, 2019.

[384] Saori Takahashi, Shin Kobayashi, and Ichiro Hiratani. Epigenetic differences between naïve and primed pluripotent stem cells. *Cellular and Molecular Life Sciences*, 75(7):1191–1203, 2018.

[385] Daniel C. Factor, Olivia Corradin, Gabriel E. Zentner, Alina Saiakhova, Lingyun Song, Josh G. Chenoweth, Ronald D. McKay, Gregory E. Crawford, Peter C. Scacheri, and Paul J. Tesar. Epigenomic comparison reveals activation of "seed" enhancers during transition from naive to primed pluripotency. *Cell Stem Cell*, 14(6):854–863, 2014.

[386] Warren A. Whyte, Steve Bilodeau, David A. Orlando, Heather A. Hoke, Garrett M. Frampton, Charles T. Foster, Shaun M. Cowley, and Richard A. Young. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature*, 482(7384):221–225, 2012.

[387] Christa Buecker, Rajini Srinivasan, Zhixiang Wu, Eliezer Calo, Dario Acampora, Tiago Faial, Antonio Simeone, Minjia Tan, Tomasz Swigut, and Joanna Wysocka. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell*, 14(6):838–853, 2014.

[388] Young Il Yeom, Guy Fuhrmann, Catherine E Ovitt, Alexander Brehm, Kazuyuki Ohbo, Michael Gross, Karin Hübner, Hans R Schöler, and Gene Expression Programme. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development*, pages 1–14, 1996.

[389] Stephanie L. Battle, Naresh Doni Jayavelu, Robert N. Azad, Jennifer Hesson, Faria N. Ahmed, Eliah G. Overbey, Joseph A. Zoller, Julie Mathieu, Hannele Ruohola-Baker, Carol B. Ware, and R. David Hawkins. Enhancer Chromatin and 3D Genome Architecture Changes from Naive to Primed Human Embryonic Stem Cell States. *Stem Cell Reports*, 12(5):1129–1144, 2019.

[390] Christina Galonska, Michael J. Ziller, Rahul Karnik, and Alexander Meissner. Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor

Binding before Global Epigenetic Reprogramming. *Cell Stem Cell*, 17(4):462–470, 2015.

[391] Giuliano Crispatzu, Rizwan Rehimi, Tomas Pachano, Tore Bleckwehl, Sara Cruz-Molina, Cally Xiao, Esther Mahabir, Hisham Bazzi, and Alvaro Rada-Iglesias. The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nature Communications*, 12(1):1–17, 2021.

[392] Tarjei S. Mikkelsen, Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae Kyung Kim, Richard P. Koche, William Lee, Eric Mendenhall, Aisling O'Donovan, Aviva Presser, Carsten Russ, Xiaohui Xie, Alexander Meissner, Marius Wernig, Rudolf Jaenisch, Chad Nusbaum, Eric S. Lander, and Bradley E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

[393] Véronique Azuara, Pascale Perry, Stephan Sauer, Mikhail Spivakov, Helle F. Jørgensen, Rosalind M. John, Mina Gouti, Miguel Casanova, Gary Warnes, Matthias Merkenschlager, and Amanda G. Fisher. Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8(5):532–538, 2006.

[394] Guangjin Pan, Shulan Tian, Jeff Nie, Chuhu Yang, Victor Ruotti, Hairong Wei, Gudrun A. Jonsdottir, Ron Stewart, and James A. Thomson. Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell*, 1(3):299–312, 2007.

[395] Xiao Dong Zhao, Xu Han, Joon Lin Chew, Jun Liu, Kuo Ping Chiu, Andre Choo, Yuriy L. Orlov, Wing Kin Sung, Atif Shahab, Vladimir A. Kuznetsov, Guillaume Bourque, Steve Oh, Yijun Ruan, Huck Hui Ng, and Chia Lin Wei. Whole-Genome Mapping of Histone H3 Lys4 and 27 Trimethylations Reveals Distinct Genomic Compartments in Human Embryonic Stem Cells. *Cell Stem Cell*, 1(3):286–298, 2007.

[396] Iying Yan, Hongshan Guo, Boqiang Hu, Rong Li, Jun Yong, Yangyu Zhao, Xu Zhi, Xiaoying Fan, Fan Guo, Xiaoye Wang, Wei Wang, Yuan Wei, Yan Wang, Lu Wen, Jie Qiao, and Fuchou Tang. Epigenomic landscape of human fetal brain, heart, and liver. *Journal of Biological Chemistry*, 291(9):4386–4398, 2016.

[397] Serap Erkek, Mizue Hisano, Ching Yeu Liang, Mark Gill, Rabih Murr, Jürgen Dieker, Dirk Schübeler, Johan Van Der Vlag, Michael B. Stadler, and Antoine H.F.M. Peters. Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nature Structural and Molecular Biology*, 20(7):868–875, 2013.

[398] Xiaoyu Liu, Chenfei Wang, Wenqiang Liu, Jingyi Li, Chong Li, Xiaochen Kou, Jiayu Chen, Yanhong Zhao, Haibo Gao, Hong Wang, Yong Zhang, Yawei Gao, and Shaorong Gao. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature*, 537(7621):558–562, 2016.

[399] Bingjie Zhang, Hui Zheng, Bo Huang, Wenzhi Li, Yunlong Xiang, Xu Peng, Jia Ming, Xiaotong Wu, Yu Zhang, Qianhua Xu, Wenqiang Liu, Xiaochen Kou, Yanhong Zhao, Wenteng He, Chong Li, Bo Chen, Yuanyuan Li, Qiujun Wang, Jing Ma, Qiangzong Yin, Kehkooi Kee, Anming Meng, Shaorong Gao, Feng Xu, Jie Na, and Wei Xie. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature*, 537(7621):553–557, 2016.

[400] Hui Zheng, Bo Huang, Bingjie Zhang, Yunlong Xiang, Zhenhai Du, Qianhua Xu, Yuanyuan Li, Qiujun Wang, Jing Ma, Xu Peng, Feng Xu, and Wei Xie. Resetting Epigenetic Memory by Reprogramming of Histone Modifications in Mammals. *Molecular Cell*, 63(6):1066–1079, 2016.

[401] Tae Young Roh, Suresh Cuddapah, Kairong Cui, and Keji Zhao. The genomic landscape of histone modifications in human T cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15782–15787, 2006.

[402] Olivia Alder, Fabrice Lavial, Anne Helness, Emily Brookes, Sandra Pinho, Anil Chandrashekran, Philippe Arnaud, Ana Pombo, Laura O'Neill, and Véronique Azuara. Ring1B and Suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment. *Development*, 137(15):2483–2492, 2010.

[403] John Arne Dahl, Andrew H. Reiner, Arne Klungland, Teruhiko Wakayama, and Philippe Collas. Histone H3 lysine 27 methylation asymmetry on developmentally-regulated promoters distinguish the first two lineages in mouse preimplantation embryos. *PLoS ONE*, 5(2), 2010.

[404] Philipp Voigt, Wee Wei Tee, and Danny Reinberg. A double take on bivalent promoters. *Genes and Development*, 27(12):1318–1338, 2013.

[405] Philipp Voigt, Gary LeRoy, William J. Drury, Barry M. Zee, Jinsook Son, David B. Beck, Nicolas L. Young, Benjamin A. Garcia, and Danny Reinberg. Asymmetrically modified nucleosomes. *Cell*, 151(1):181–193, 2012.

[406] Deqing Hu, Alexander S. Garruss, Xin Gao, Marc A. Morgan, Malcolm Cook, Edwin R. Smith, and Ali Shilatifard. The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nature Structural and Molecular Biology*, 20(9):1093–1097, 2013.

[407] Glòria Mas, Enrique Blanco, Cecilia Ballaré, Miriam Sansó, Yannick G. Spill, Deqing Hu, Yuki Aoi, François Le Dily, Ali Shilatifard, Marc A. Marti-Renom, and Luciano Di Croce. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nature Genetics*, 50(10):1452–1462, 2018.

[408] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[409] Aaron D. Goldberg, Laura A. Banaszynski, Kyung Min Noh, Peter W. Lewis, Simon J. Elsaesser, Sonja Stadler, Scott Dewell, Martin Law, Xingyi Guo, Xuan Li, Duancheng Wen, Ariane Chapgier, Russell C. DeKelver, Jeffrey C. Miller, Ya Li Lee, Elizabeth A. Boydston, Michael C. Holmes, Philip D. Gregory, John M. Greally, Shahin Rafii, Chingwen Yang, Peter J. Scambler, David Garrick, Richard J. Gibbons, Douglas R. Higgs, Ileana M. Cristea, Fyodor D. Urnov, Deyou Zheng, and C. David Allis. Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions. *Cell*, 140(5):678–691, 2010.

[410] Manching Ku, Jacob D. Jaffe, Richard P. Koche, Esther Rheinbay, Mitsuhiro Endoh, Haruhiko Koseki, Steven A. Carr, and Bradley E. Bernstein. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biology*, 13(10), 2012.

[411] Gangqing Hu, Kairong Cui, Daniel Northrup, Chengyu Liu, Chaochen Wang, Qingsong Tang, Kai Ge, David Levens, Colyn Crane-Robinson, and Keji Zhao. H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell*, 12(2):180–192, 2013.

[412] Magnus D. Lynch, Andrew J.H. Smith, Marco De Gobbi, Maria Flenley, Jim R. Hughes, Douglas Vernimmen, Helena Ayyub, Jacqueline A. Sharpe, Jacqueline A. Sloane-Stanley, Linda Sutherland, Stephen Meek, Tom Burdon, Richard J. Gibbons, David Garrick, and Douglas R. Higgs. An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO Journal*, 31(2):317–329, 2012.

[413] Mélanie Eckersley-Maslin, Celia Alda-Catalinas, Marloes Blotenburg, Elisa Kreibich, Christel Krueger, and Wolf Reik. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes and Development*, 33(3-4):194–208, 2019.

[414] Mélanie A. Eckersley-Maslin, Aled Parry, Marloes Blotenburg, Christel Krueger, Yoko Ito, Valar Nila Roamio Franklin, Masashi Narita, Clive S. D'Santos, and Wolf Reik. Epigenetic priming by Dppa2 and 4 in pluripotency facilitates multi-lineage commitment. *Nature Structural and Molecular Biology*, 27(8):696–705, 2020.

[415] Rohan N. Shah, Adrian T. Grzybowski, Jimmy Elias, Zhonglei Chen, Takamitsu Hattori, Carolin C. Lechner, Peter W. Lewis, Shohei Koide, Beat Fierz, and Alexander J. Ruthenburg. Re-evaluating the role of nucleosomal bivalency in early development. *bioRxiv*, page 2021.09.09.458948, 2021.

[416] Andrew D. King, Kevin Huang, Liudmilla Rubbi, Shuo Liu, Cun Yu Wang, Yinsheng Wang, Matteo Pellegrini, and Guoping Fan. Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells. *Cell Reports*, 17(1):289–302, 2016.

[417] Stephan H. Bernhart, Helene Kretzmer, Lesca M. Holdt, Frank Jühling, Ole Ammerpohl, Anke K. Bergmann, Bernd H. Northoff, Gero Doose, Reiner Siebert, Peter F. Stadler, and Steve Hoffmann. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Scientific Reports*, 6:1–18, 2016.

[418] Sayyed K. Zaidi, Seth E. Frietze, Jonathan A. Gordon, Jessica L. Heath, Terri Messier, Deli Hong, Joseph R. Boyd, Mingu Kang, Anthony N. Imbalzano, Jane B. Lian, Janet L. Stein, and Gary S. Stein. Bivalent Epigenetic Control of Oncofetal Gene Expression in Cancer. *Molecular and Cellular Biology*, 37(23), 2017.

[419] Jiabiao Hu, Yong Lei, Wing-Ki Wong, Senquan Liu, Kai-Chuen Lee, Xiangjun He, Wenxing You, Rui Zhou, Jun-Tao Guo, Xiongfong Chen, Xianlu Peng, Hao Sun, He Huang, Hui Zhao, and Bo Feng. Direct activation of human and mouse Oct4 genes using engineered TALE and Cas9 transcription factors. *Nucleic Acids Research*, 42(7):4375–4390, 2014.

[420] Benjamin P. Kleinstiver, Michelle S. Prew, Shengdar Q. Tsai, Ved V. Topkar, Nhu T. Nguyen, Zongli Zheng, Andrew P.W. Gonzales, Zhuyun Li, Randall T. Peterson, Jing Ruey Joanna Yeh, Martin J. Aryee, and J. Keith Joung. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 523(7561):481–485, 2015.

[421] Lei S. Qi, Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

[422] Albert W. Cheng, Haoyi Wang, Hui Yang, Linyu Shi, Yarden Katz, Thorold W. Theunissen, Sudharshan Rangarajan, Chikdu S. Shivalila, Daniel B. Dadon, and Rudolf Jaenisch. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Research*, 23(10):1163–1171, 2013.

[423] Ruilin Tian, Anthony Abarientos, Jason Hong, Sayed Hadi Hashemi, Rui Yan, Nina Dräger, Kun Leng, Mike A. Nalls, Andrew B. Singleton, Ke Xu, Faraz Faghri, and Martin Kampmann. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24(7):1020–1034, 2021.

[424] Maria Rostovskaya, Giuliano G. Stirparo, and Austin Smith. Capacitation of human naïve pluripotent stem cells for multi-lineage differentiation. *Development (Cambridge)*, 146(7), 2019.

[425] Hatice S. Kaya-Okur, Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1–10, 2019.

[426] Charles P. Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, and Eric S Lander. Sys-

tematic mapping of functional enhancer-promoter connections with CRISPR interference. 354(6313):769–773, 2017.

[427] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[428] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Shirley. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), 2008.

[429] Thomas S. Carroll, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in Genetics*, 5(APR):1–11, 2014.

[430] Jason Ernst and Manolis Kellis. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012.

[431] Steven Wingett, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Research*, 4:1–12, 2015.

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

[433] Paula Freire-Pritchett, Helen Ray-Jones, Monica Della Rosa, Chris Q. Eijsbouts, William R. Orchard, Steven W. Wingett, Chris Wallace, Jonathan Cairns, Mikhail Spivakov, and Valeriya Malysheva. *Detecting chromosomal interactions in Capture Hi-C data with CHiCAGO and companion tools*, volume 16. Springer US, 2021.

[434] A Kassambara and F Mundt. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

[435] R Kolde. Implementation of heatmaps that offers more control over dimensions and appearance.

[436] Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, and Cristina Sisu. Gencode 2021. *Nucleic Acids Research*, 49(D1):D916–D923, 2021.

[437] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[438] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.

[439] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

[440] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

[441] John Fox. Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8:1–27, 2003.

[442] H Pagès. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.

[443] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2018.

[444] Helge G. Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

[445] A Droit, R Gottardo, G Robertson, and L Li. rGADEM: de novo motif discovery.

[446] Leping Li. GADEM: A genetic algorithm guided formation of spaced dyads coupled with an em algorithm for motif discovery. *Journal of Computational Biology*, 16(2):317–329, 2009.

[447] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and K Christopher. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. 38(4):576–589, 2011.

[448] Jairo Navarro Gonzalez, Ann S. Zweig, Matthew L. Speir, Daniel Schmelter, Kate R. Rosenbloom, Brian J. Raney, Conner C. Powell, Luis R. Nassar, Nathan D. Maulding, Christopher M. Lee, Brian T. Lee, Angie S. Hinrichs, Alastair C. Fyfe, Jason D. Fernandes, Mark Diekhans, Hiram Clawson, Jonathan Casper, Anna Benet-Pagès, Galt P. Barber, David Haussler, Robert M. Kuhn, Maximilian Haeussler, and W. James Kent. The UCSC genome browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, 2021.

[449] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):1–10, 2013.

[450] Bernat Gel, Anna Díez-Villanueva, Eduard Serra, Marcus Buschbeck, Miguel A. Peinado, and Roberto Malinverni. RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, 32(2):289–291, 2016.

[451] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[452] Valeriya Malysheva, Stefan Schoenfelder, Mikhail Spivakov, Takashi Nagano, and Peter Fraser. WO2021064430A1, 2021.

[453] Simone Picelli, Asa K. Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, 24(12):2033–2040, 2014.

[454] Xingqi Chen, Ying Shen, Will Draper, Jason D. Buenrostro, Ulrike Litzenburger, Seung Woo Cho, Ansuman T. Satpathy, Ava C. Carter, Rajarshi P. Ghosh, Alexandra East-Seletsky, Jennifer A. Doudna, William J. Greenleaf, Jan T. Liphardt, and

Howard Y. Chang. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature Methods*, 13(12):1013–1020, 2016.

[455] Jie Huang, Yongpeng Jiang, Haonan Zheng, and Xiong Ji. BAT Hi-C maps global chromatin interactions in an efficient and economical way. *Methods*, 170(March 2019):38–47, 2020.

[456] S. A. McCommas and M. Syvanen. Temporal control of transposition in Tn5. *Journal of bacteriology*, 170(2):889–894, 1988.

[457] Ryota Suganuma, Pawel Pelczar, Jean François Spetz, Barbara Hohn, Ryuzo Yanagimachi, and Stefan Moisyadi. Tn5 transposase-mediated mouse transgenesis. *Biology of Reproduction*, 73(6):1157–1163, 2005.

[458] Chongyi Chen, Dong Xing, Longzhi Tan, Heng Li, Guangyu Zhou, Lei Huang, and X. Sunney Xie. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science*, 356(6334):189–194, 2017.

[459] Chun Su, Matthew C. Pahl, Struan F.A. Grant, and Andrew D. Wells. Restriction enzyme selection dictates detection range sensitivity in chromatin conformation capture-based variant-to-gene mapping approaches. *Human Genetics*, 140(10):1441–1448, 2021.

[460] Linden Disney-hogg, Ben Kinnersley, and Richard Houlston. Algorithmic considerations when analysing capture Hi-C data [version 1 ; peer review : 1 approved , 1 approved with reservations]. 2022.

[461] Carrie A. Davis, Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, Kriti Jain, Ulugbek K. Baymuradov, Aditi K. Narayanan, Kathrina C. Onate, Keenan Graham, Stuart R. Miyasato, Timothy R. Dreszer, J. Seth Strattan, Otto Jolanki, Forrest Y. Tanaka, and J. Michael Cherry. The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018.

[462] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative Genome Viewer. *Nature Biotechnology*, 29(1):24–6, 2011.

[463] Alberto De Iaco, Alexandre Coudray, Julien Duc, and Didier Trono. DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO reports*, 20(5):1–10, 2019.

[464] Yao Long Yan, Chao Zhang, Jing Hao, Xue Lian Wang, Jia Ming, Li Mi, Jie Na, Xinli Hu, and Yangming Wang. *DPPA2/4 and SUMO E3 ligase PIAS4 opposingly regulate zygotic transcriptional program*, volume 17. 2019.

[465] Charles Hernandez, Zheng Wang, Bulat Ramazanov, Yin Tang, Sameet Mehta, Cheryl Dambrot, Yu Wei Lee, Kaleab Tessema, Ishan Kumar, Michael Astudillo, Thomas A. Neubert, Shangqin Guo, and Natalia B. Ivanova. Dppa2/4 Facilitate Epigenetic Remodeling during Reprogramming to Pluripotency. *Cell Stem Cell*, 23(3):396–411, 2018.

[466] Erik Engelen, Johannes H. Brandsma, Maaike J. Moen, Luca Signorile, Dick H.W. Dekkers, Jeroen Demmers, Christel E.M. Kockx, Zehila Ozgür, Wilfred F.J. Van Ijcken, Debbie L.C. Van Den Berg, and Raymond A. Poot. Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. *Nature Communications*, 6(May), 2015.

[467] C.P. Fulco, J. Nasser, T.R. Jones, G. Munson, D.T. Bergman, V. Subramanian, S.R. Grossman, R. Anyoha, B.R. Doughty, T.A. Patwardhan, T.H. Nguyen, M. Kane, E.M. Perez, N.C. Durand, C.A. Lareau, E.K. Stamenova, E.L. Aiden, E.S. Lander, and J.M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12), 2019.

[468] Stefan Schoenfelder, Robert Sugar, Andrew Dimond, Biola Maria Javierre, Harry Armstrong, Borbala Mifsud, Emilia Dimitrova, Louise Matheson, Filipe Tavares-Cadete, Mayra Furlan-Magaril, Anne Segonds-Pichon, Wiktor Jurkowski, Steven W. Wingett, Kristina Tabbada, Simon Andrews, Bram Herman, Emily Leproust, Cameron S. Osborne, Haruhiko Koseki, Peter Fraser, Nicholas M. Luscombe, and Sarah Elderkin. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature Genetics*, 47(10):1179–1186, 2015.

[469] Manching Ku, Richard P. Koche, Esther Rheinbay, Eric M. Mendenhall, Mitsuhiro Endoh, Tarjei S. Mikkelsen, Aviva Presser, Chad Nusbaum, Xiaohui Xie, Andrew S. Chi, Mazhar Adli, Simon Kasif, Leon M. Ptaszek, Chad A. Cowan, Eric S. Lander,

Haruhiko Koseki, and Bradley E. Bernstein. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genetics*, 4(10), 2008.

[470] Emilia Dimitrova, Angelika Feldmann, Robin H van der Weide, Koen D Flach, Anna Lastuvkova, Elzo de Wit, and Robert J Klose. Distinct roles for CDK-Mediator in controlling Polycomb-dependent chromosomal interactions and priming genes for induction. *bioRxiv*, 1:2021.11.04.467119, 2021.

[471] Donnchadh S. Dunican, Heidi K. Mjoseng, Leanne Duthie, Ilya M. Flyamer, Wendy A. Bickmore, and Richard R. Meehan. Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. *BMC Biology*, 18(1):1–21, 2020.

[472] Michal Levo, João Raimundo, Xin Yang Bing, Zachary Sisco, Philippe J Batut, Sergey Ryabichko, Thomas Gregor, and Michael S Levine. Transcriptional coupling of distant regulatory genes in living embryos. (March 2021), 2022.

[473] Licia Selleri, Vincenzo Zappavigna, and Elisabetta Ferretti. 'Building a perfect body': Control of vertebrate organogenesis by PBX-dependent regulatory networks. *Genes and Development*, 33(5-6):258–275, 2019.

[474] Nikhil Gupta, Lounis Yakhou, Julien Richard Albert, Fumihito Miura, Laure Ferry, Olivier Kirsh, Marthe Laisné, Kosuke Yamaguchi, Cécilia Domrane, Frédéric Bonhomme, Arpita Sarkar, Marine Delagrange, Bertrand Ducos, Maxim V C Greenberg, Gael Cristofari, Sebastian Bultmann, Takashi Ito, and Pierre-Antoine Defossez. A genome-wide knock-out screen for actors of epigenetic silencing reveals new regulators of germline genes and 2-cell like cell state. *bioRxiv*, page 2021.05.03.442415, 2021.

[475] V. Loubiere, G. L. Papadopoulos, Q. Szabo, A. M. Martinez, and G. Cavalli. Widespread activation of developmental gene expression characterized by PRC1-dependent chromatin looping. *Science Advances*, 6(2), 2020.

[476] Giorgio Oliviero, Nayla Munawar, Ariane Watson, Gundula Streubel, Gwendolyn Manning, Vivian Bardwell, Adrian P. Bracken, and Gerard Cagney. The variant Polycomb Repressor Complex 1 component PCGF1 interacts with a pluripotency

sub-network that includes DPPA4, a regulator of embryogenesis. *Scientific Reports*, 5(August):1–11, 2015.

[477] Riccardo Arrigucci, Yuri Bushkin, Felix Radford, Karim Lakehal, Pooja Vir, Richard Pine, December Martin, Jeffrey Sugarman, Yanlin Zhao, George S. Yap, Alfred A. Lardizabal, Sanjay Tyagi, and Maria Laura Gennaro. FISH-Flow, a protocol for the concurrent detection of mRNA and protein in single cells using fluorescence in situ hybridization and flow cytometry. *Nature Protocols*, 2017.

[478] Niannian Li, Kairang Jin, Yanmin Bai, Haifeng Fu, Lin Liu, and Bin Liu. Tn5 transposase applied in genomics research. *International Journal of Molecular Sciences*, 21(21):1–15, 2020.

[479] Chew Yee Ngan, Chee Hong Wong, Harianto Tjong, Wenbo Wang, Rachel L. Goldfeder, Cindy Choi, Hao He, Liang Gong, Junyan Lin, Barbara Urban, Julianna Chow, Meihong Li, Joanne Lim, Vivek Philip, Stephen A. Murray, Haoyi Wang, and Chia Lin Wei. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nature Genetics*, 52(3):264–272, 2020.

[480] Katy McLaughlin, Ilya M. Flyamer, John P. Thomson, Heidi K. Mjoseng, Ruchi Shukla, Iain Williamson, Graeme R. Grimes, Robert S. Illingworth, Ian R. Adams, Sari Pennings, Richard R. Meehan, and Wendy A. Bickmore. DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. *Cell Reports*, 29(7):1974–1985, 2019.

[481] Dhirendra Kumar, Senthilkumar Cinghu, Andrew J. Oldfield, Pengyi Yang, and Raja Jothi. Decoding the function of bivalent chromatin in development and cancer. *Genome Research*, 31(12):2170–2184, 2021.

[482] Takashi Kondo, Shinsuke Ito, and Haruhiko Koseki. Polycomb in Transcriptional Phase Transition of Developmental Genes. *Trends in Biochemical Sciences*, 41(1):9–19, 2016.

[483] Kyoichi Isono, Takaho A. Endo, Manching Ku, Daisuke Yamada, Rie Suzuki, Jafar Sharif, Tomoyuki Ishikura, Tetsuro Toyoda, Bradley E. Bernstein, and Haruhiko Koseki. SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. *Developmental Cell*, 26(6):565–577, 2013.

[484] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–472, 2012.

[485] Matthew Denholtz, Giancarlo Bonora, Constantinos Chronis, Erik Splinter, Wouter de Laat, Jason Ernst, Matteo Pellegrini, and Kathrin Plath. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and Polycomb proteins in genome organization. *Cell Stem Cell*, 13(5), 2013.

[486] Banushree Kumar, Carmen Navarro, Nerges Winblad, John P. Schell, Cheng Zhao, Jere Weltner, Laura Baqué-Vidal, Angelo Salazar Mantero, Sophie Petropoulos, Fredrik Lanner, and Simon J. Elsässer. Polycomb repressive complex 2 shields naïve human pluripotent cells from trophectoderm differentiation. *Nature Cell Biology*, 2022.

[487] Dick W Zijlmans, Irene Talon, Sigrid Verhelst, Adam Bendall, Karlien Van Nerum, Alok Javali, Andrew A Malcolm, Sam S F A Van Knippenberg, Laura Biggins, San Kit To, Adrian Janiszewski, Danielle Admiraal, Ruth Knops, Nikky Corthout, Bradley P Balaton, Grigorios Georgolopoulos, Amitesh Panda, Natarajan V Bhanu, Amanda J Collier, Charlene Fabian, Ryan N Allsop, Joel Chappell, Thi Xuan, Ai Pham, Michael Oberhuemer, Cankat Ertekin, Lotte Vanheer, Paraskevi Athanasouli, Frederic Lluis, Dieter Deforce, Joop H Jansen, Benjamin A Garcia, Michiel Vermeulen, Nicolas Rivron, Maarten Dhaenens, Hendrik Marks, Rugg-Gunn Peter J., Pasque, and Vincent. Integrated multi-omics reveal polycomb repressive complex 2 restricts human trophoblast induction. *Nature Cell Biology*, 2022.

[488] Sergi Aranda, Livia Condemi, and Luciano Di Croce. PRC2 shields the potency of human stem cells. *Nature Cell Biology*, 2022.

[489] Monica Della Rosa and Mikhail Spivakov. Silencers in the spotlight. *Nature Genetics*, 52(3):244–245, 2020.

[490] Shiri Levy, Logeshwaran Somasundaram, Infencia Xavier Raj, Diego Ic-Mex, Ashish Phal, Sven Schmidt, Weng I. Ng, Daniel Mar, Justin Decarreau, Nicholas Moss, Ammar Alghadeer, Henrik Honkanen, Jay Sarthy, Nicholas Vitanza, R. David Hawkins, Julie Mathieu, Yuliang Wang, David Baker, Karol Bomsztyk,

and Hannele Ruohola-Baker. dCas9 fusion to computer-designed PRC2 inhibitor reveals functional TATA box in distal promoter region. *Cell Reports*, 38(9):110457, 2022.

A Supplementary Figures

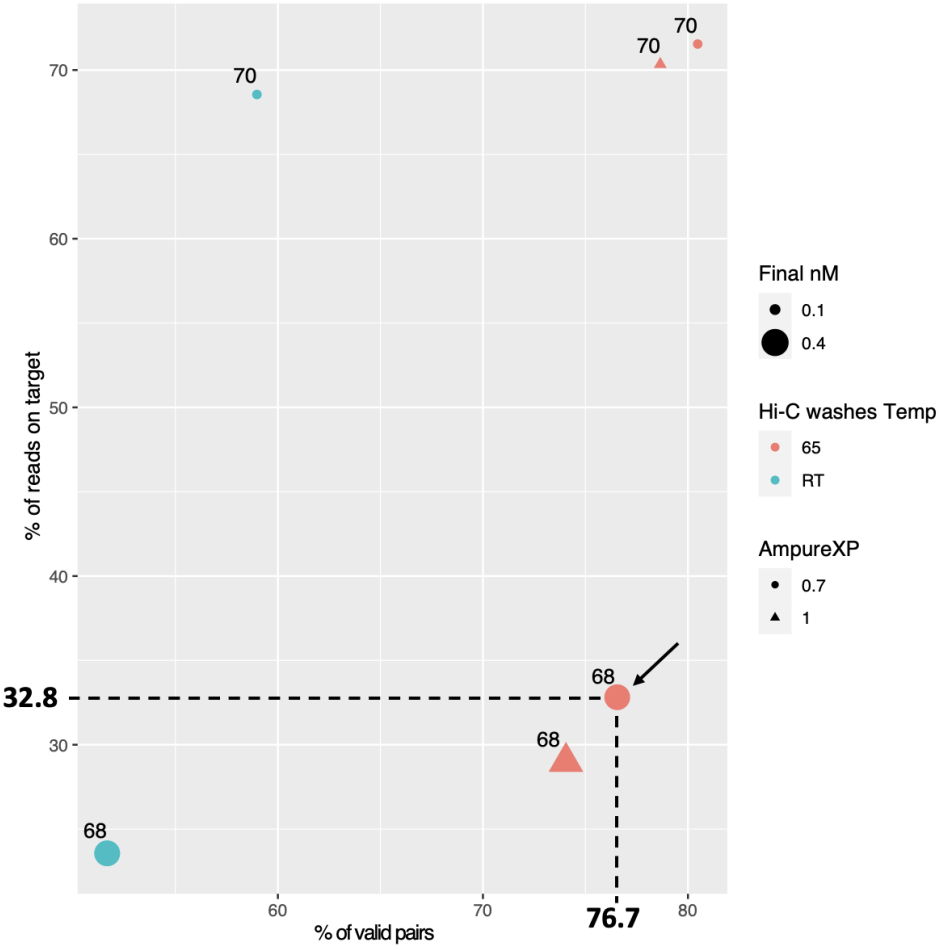


Figure A.1: **Identification of the optimal conditions for the generation of CHi-C libraries.** Scatter plot summarizing the optimization conditions an their effect on percentage of reads on target (y-axis) and percentage of valid pairs (x-axis). The chosen final conditions established are highlighted by the black arrow, aiming to get the optimal combination between percentage of valid pairs, percentage of reads on target (y-axis) and the final library yield obtained of the PECHi-C library (final concentration, expressed in nano molar or nM, is represented by the size of the points on the scatter plot). Both percentage of reads on target and percentage of valid pairs have been averaged across four different samples for each condition.

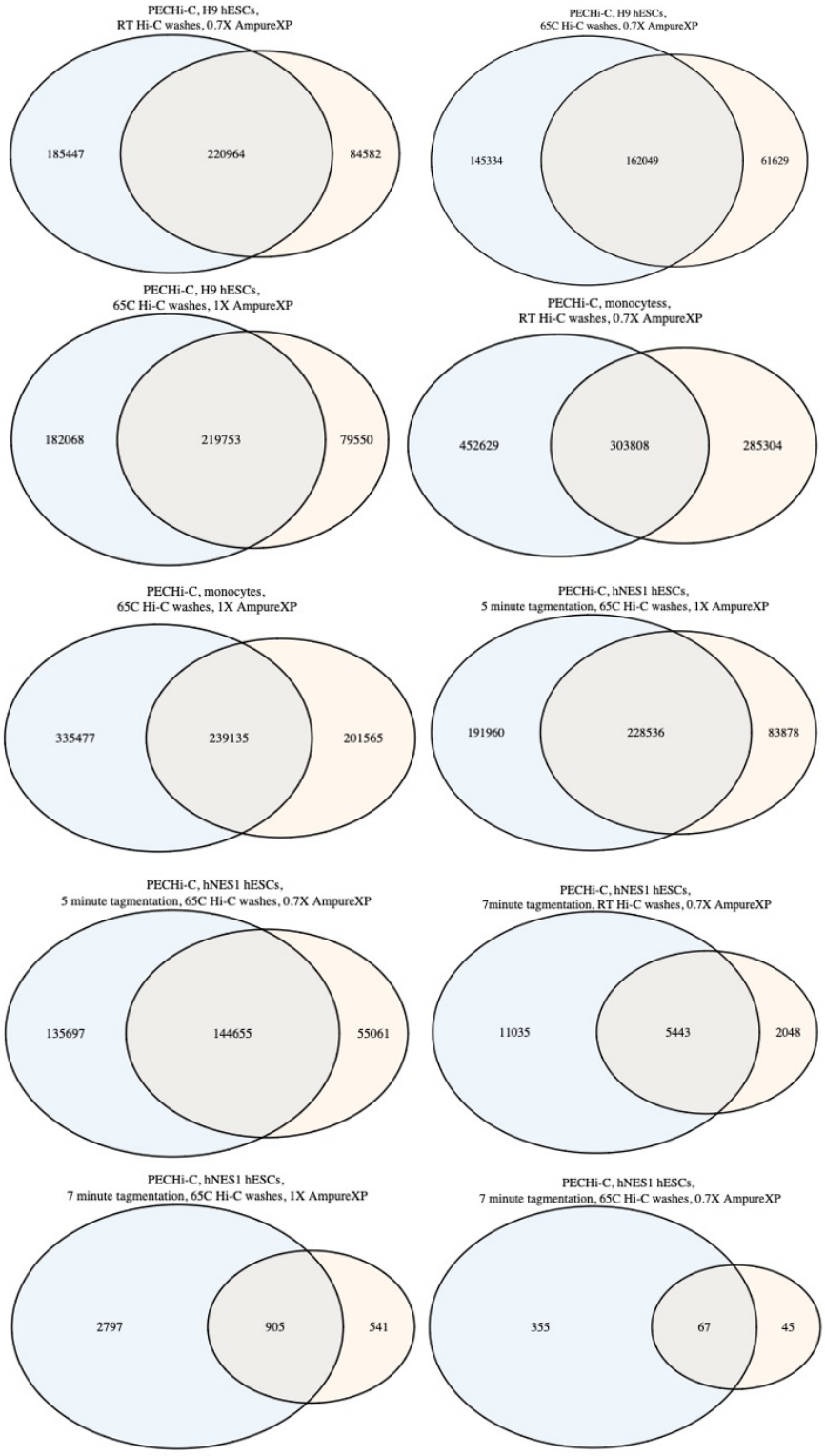


Figure A.2: **Overlap of CHiCAGO detected interactions following 20% data down-sampling.** Venn diagrams showing the overlap of CHiCAGO detected interactions in CHi-C data generated in primed (H9) hESCs, monocytes and naïve (hNES1) hESCs (light blue) at the different optimization conditions following a 20% down-sampling of the CHi-C data (light orange). Absolute number of overlapping interactions and sample-specific interactions is given within each Venn diagram. Overlapping percentage ranged between 75% and 51% for all samples, following a 20% down-sampling.

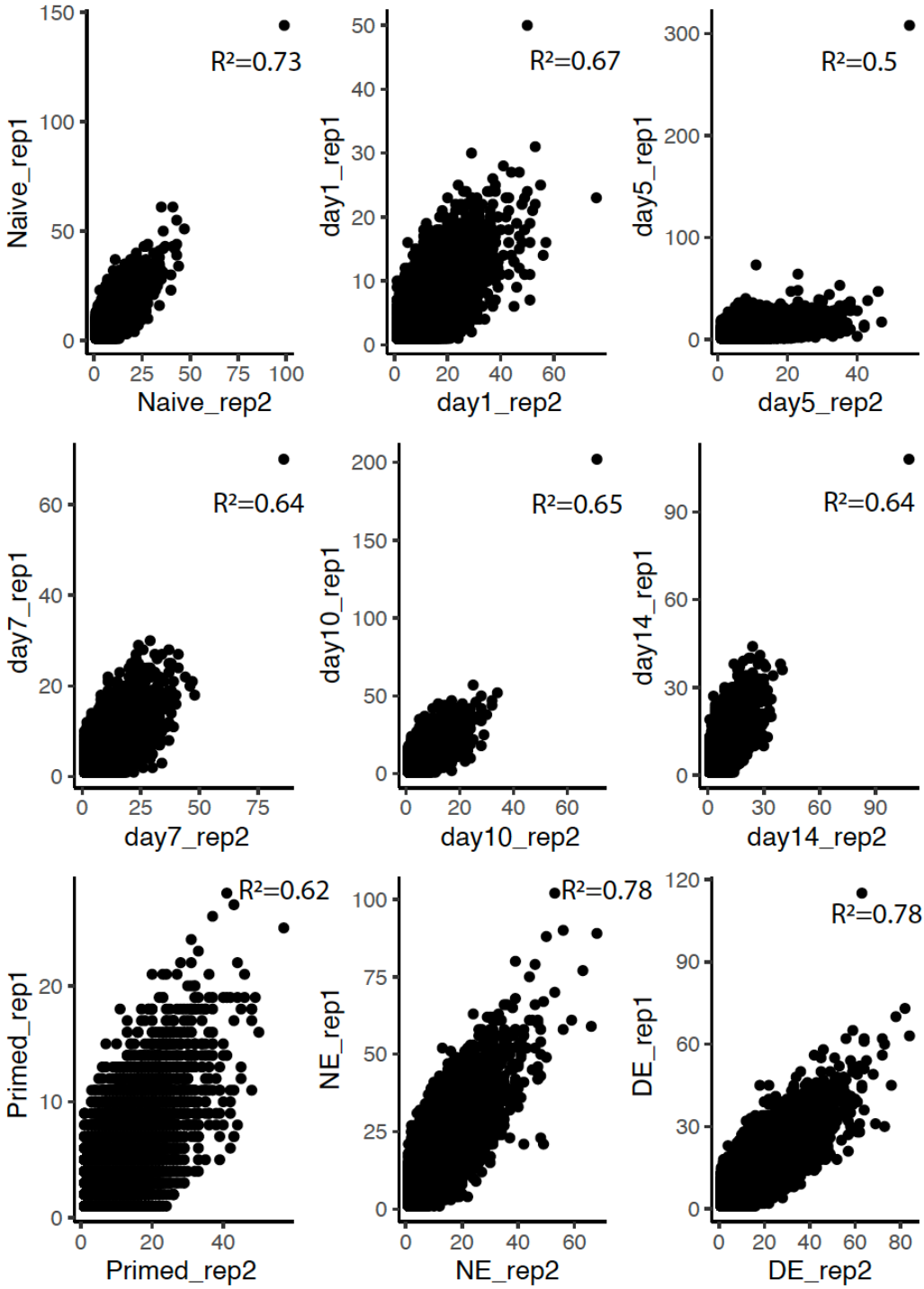


Figure A.3: **Biological replicates correlation of read counts per interaction pair detected by CHiCAGO.** Scatter plots show Pearson's correlation (Pearson's correlation coefficient for each replicate pair shown on the scatter plot) of contact pairs detected by CHiCAGO between biological replicates for each time point as hESCs transitioned between the naïve(hNES1) and primed (H9) state of pluripotency. On the x-axis CHiCAGO normalized read counts of biological replicate 2 (rep2) for each time point. On the y-axis CHiCAGO normalized read counts of biological replicate 1 (rep1) for each time point.

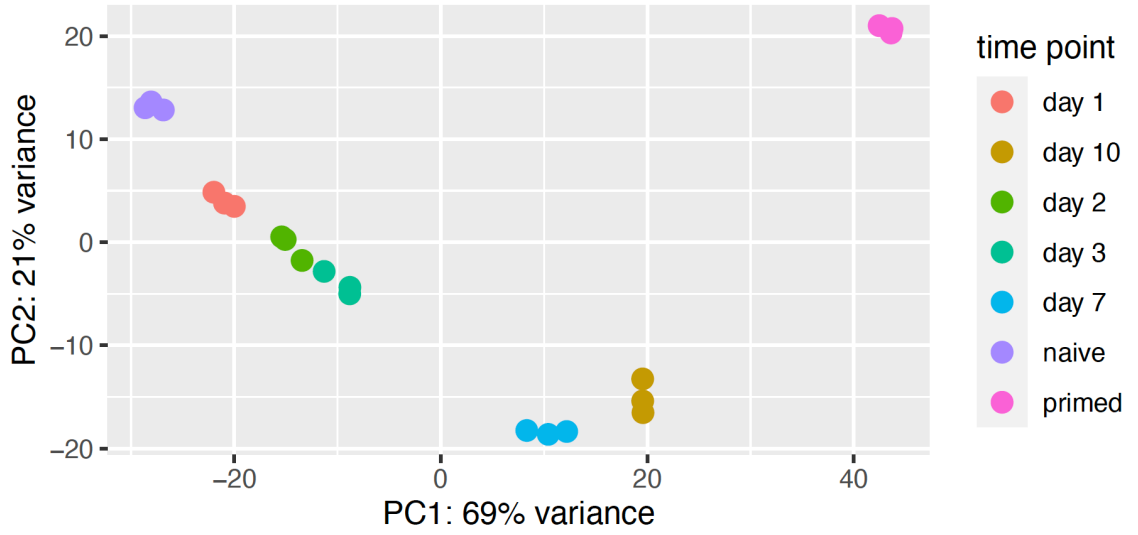


Figure A.4: **PCA analysis of gene expression changes of individual timepoints upon the naïve-to-primed transition in hESCs.** PCA analysis of previously published RNA-seq data GSE123055 generated for hNES1 hESCs during the time course of the naïve-to-primed (primed H9 hESCs) transition [424]. The PCA plot shows a clear dependency of differential gene expression on time which can be appreciated already at the very early stages of the transition, as shown by the left side of the PCA plot, along Dimension 1 (x-axis), explaining 69% of the variability.

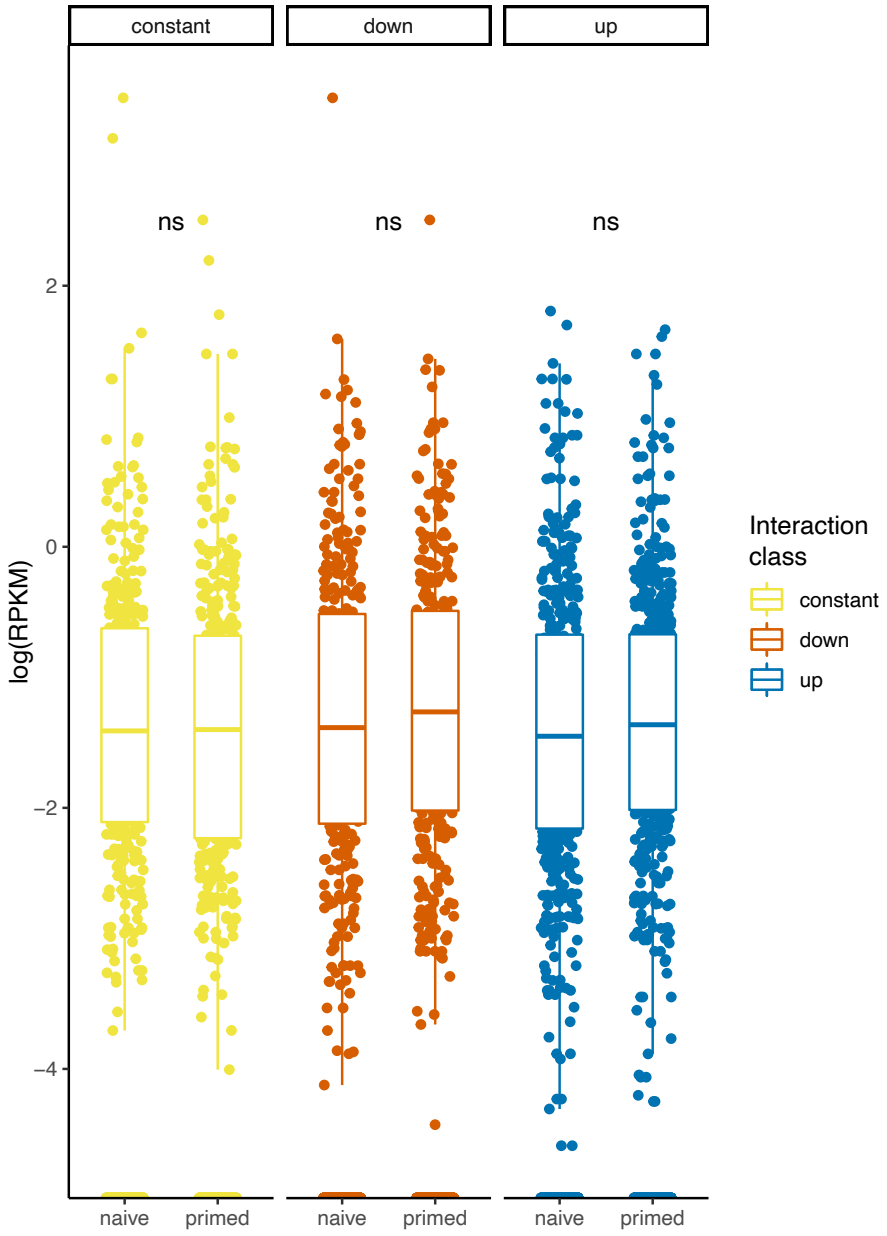


Figure A.5: **DNA accessibility of PEs within different interaction classes in naïve (hNES1) and primed (H9) hESCs.** Box plots showing no significant difference (Wilcoxon test p-value>0.05, ns) of DNA accessibility levels (assessed via ATACseq profiling) of PEs across the different interaction classes in naïve (hNES1) and primed (H9) hESCs (RPKM of ATACseq read counts on a logarithmic scale is represented on the y-axis).

B sgRNAs sequences

Gene	Target	Sequence (5′ - 3′)
NEUROD1	promoter	GGTCCGCGGAGTCTCTAAC
CXCR4	promoter	GGGGCAGACGCGAGGAAGGA GGGACCCTGCTGTTTGCGGG GGCCTCTGGGAGGTCCTGTC GCTAGGAACGCGTCTCTCTG GAAAGCGCGGGGAATGGCGT
GATA1	promoter	CTGAGCTTGCCACATCCCCA
GATA1	enhancer	CCATGGGGCCTGGACCAAGC GGCCTGACGGAGAAGACGCG GGGAAGGCTTCCGAGAAGAG GACGGAGAAGACGCGCGGCC GTCTCCCCCAAAGCCTGATC
Scramble sgRNA	N/A	AAGATGAAAGGAAAGGCGTT

C Sanger Sequencing and RT-qPCR primers

Target	Forward (5′ - 3′)	Reverse (5′ - 3′)
hU6	GAGGGCCTATTTCCCATGATT	N/A
NEUROD1	AGACTATCACTGCTCAGGACCT ACTATCACTGCTCAGGACCTACT ATCACTGCTCAGGACCTACT	GGAGCCAATGATTATGCCACC GGAGCCAATGATTATGCCACC AGTTCTCAGTCCTGGTGTTCC
CXCR4	TTACCATGGAGGGGATCAGT	ATAGTCCCCTGAGCCCATTT
GATA1	CTACTACAGGGACGCTGAGG	CCCCTCCTACAGTTGAGCAA
PRDM1	CTACCCTTATCCCGGAGAGC CCCTTATCCCGGAGAGCTGA	CGGTAGAGGTCCTTTCCTTTG GCTCGGTTGCTTTAGACTGC
ARX	CCACGTTCACCAGCTACCAG TTCCAGAAGACGCACTACCC TTCCAGAAGACGCACTACCC	CCTGCCTTCTCCCGCTTG GGAGGTAGGCTCGGGAAGG CGGTCAAGTCCAGCCTCATG
LHX6	CGTCTGCAGGCAAGAACATC ACGCCATCTGTCTGCTCAC AGGCAAGAACATCTGCTCCA	GCTGGCGTAGATCTGTCGG GCACCTTCTCCTCGACCAG CACGTGCCAGATGAGGTTGT
GAPDH	CAATGACCCCTTCATTGACC	TGGGTGGAATCATATTGGAA

D Capture Hi-C oligos & primers

Mosaic End double-stranded (MEDS)

MED name	Sequence (5′ - 3′)
ME-Rev	[phos]**CTGTCTCTTATACACATCT**
1030 (A):	GTCTCGTGGGCTCGG**AGATGTGTATAAGAGACAG**
1031 (B):	TCGTCGGCAGCGTC**AGATGTGTATAAGAGACAG**

CUSTOM BLOCKERS

Blocker	Sequence (5′ - 3′)
P5-FCA-F	GTGTAGATCTCGGTGGTCGCCGTATCATT
i5-F	CTGTCTCTTATACACATCTGACGCTGCCGACGA
i7-R	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
FCA-P7-R	ATCTCGTATGCCGTCTTCTGCTTG
P5-FCA-R*	AATGATACGGCGACCACCGAGATCTACAC
i5-Rdd	TCGTCGGCAGCGTCAGATGTGTATAAGAGA/3ddC/
dd-i7F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGA/3ddC/
FCA-P7F*	CAAGCAGAAGACGGCATACGAGAT

* these primers are used for final Capture Hi-C library PCR amplification

SEQUENCING BARCODES (i7/i5)

Name	Barcode (5′ - 3′)
N701	TAAGGCGA
N702	CGTACTAG
N703	AGGCAGAA
N704	TCCTGAGC
N705	GGACTCCT
N706	TAGGCATG
N707	CTCTCTAC
N708	CAGAGAGG
N709	GCTACGCT
N710	CGAGGCTG
N711	AAGAGGCA
N712	GTAGAGGA
S501	TAGATCGC
S502	CTCTCTAT
S503	TATCCTCT
S504	AGAGTAGA
S505	GTAAGGAG
S506	ACTGCATA
S507	AAGGAGTA
S508	CTAAGCCT

E TRAP analysis hits

TF	p.value
AP2A	0.00332131
AP2B	0.00038131
AP2C	0.00219611
BACH1	0.04495962
BC11A	0.04686373
CDX1	0.00170258
CDX2	0.00379179
CEBPG	0.02812827
CREM	0.02697879
DLX3	0.01411492
DUX4	0.0010966
E2F2	0.00568309
E2F4	0.00412359
ESR2	0.01231609
EVI1	0.01697184
FOXC1	0.0032927
FOXK1	0.02176862
FOXO4	0.04391937
FOXQ1	0.00279279
GATA6	0.02268605

HNF1A	0.04341883
HNF6	0.00145787
HXA10	0.01157716
HXA13	0.00801905
HXA1	5.95E-05
HXA9	0.02104755
HXB13	0.0055781
HXB7	0.00180799
HXB8	0.00285006
HXC9	0.00297195
INSM1	0.01447047
IRF1	0.03395293
IRF8	0.04574392
ITF2	0.01652261
KAISO	0.02756117
KLF12	0.00029775
KLF1	3.03E-06
KLF3	0.00856439
KLF4	0.02869557
KLF6	0.01673713
KLF9	0.00607187
LEF1	0.04462963
LHX2	0.02418709
LHX3	0.01622611
MAFB	0.01604498
MAFG	0.00807159
MAFK	0.00197512
MAF	0.04438212
MBD2	0.00408458
MECP2	0.00644276

MEF2D	0.04794223
MEIS1	0.0261279
MTF1	0.0141626
MYF6	0.02967572
NF2L1	0.00401926
NFAC1	0.02135473
NFAC2	0.02772681
NFIC	0.02146264
NFKB1	0.03883429
NOBOX	0.00184588
NR1H4	0.00440965
NR1I2	0.01528952
NR1I3	0.0314311
NR2C2	0.01381371
NR4A1	0.04434837
NRF1	0.01493866
OZF	0.00413778
P53	0.01067134
P73	0.0172954
PAX6	0.0173729
PBX2	5.55E-05
PDX1	0.02033323
PIT1	0.0009407
PRRX2	0.01220784
RFX1	0.00099143
RXRG	0.01453472
SMCA1	0.00573454
SOX3	0.01959845
SOX5	0.00224022
SP1	9.95E-05
SP2	0.00034657

SP3	0.00382615
SP4	0.01738932
SRBP2	0.02889535
SRY	0.04043569
STAT4	0.02768262
TBP	0.00543033
TCF7	0.01624817
THAP1	0.00015314
TWST1	0.04679591
USF2	0.04751776
Z354A	0.00372935
ZBT14	0.04058786
ZBT48	0.02503352
ZBT7A	0.00488493
ZEB1	0.04433819
ZFP82	0.0230583
ZFX	0.03771024
ZIC1	0.01767118
ZN121	0.01168984
ZN136	0.02182717
ZN143	0.01560394
ZN274	0.00913162
ZN317	0.01388062
ZN320	0.01666137
ZN329	0.00527297
ZN449	0.01713126
ZN554	0.02546024
ZN563	0.02660629
ZN667	0.00655698
ZN680	0.00269107
ZNF76	0.0332219
ZSC22	0.00811864