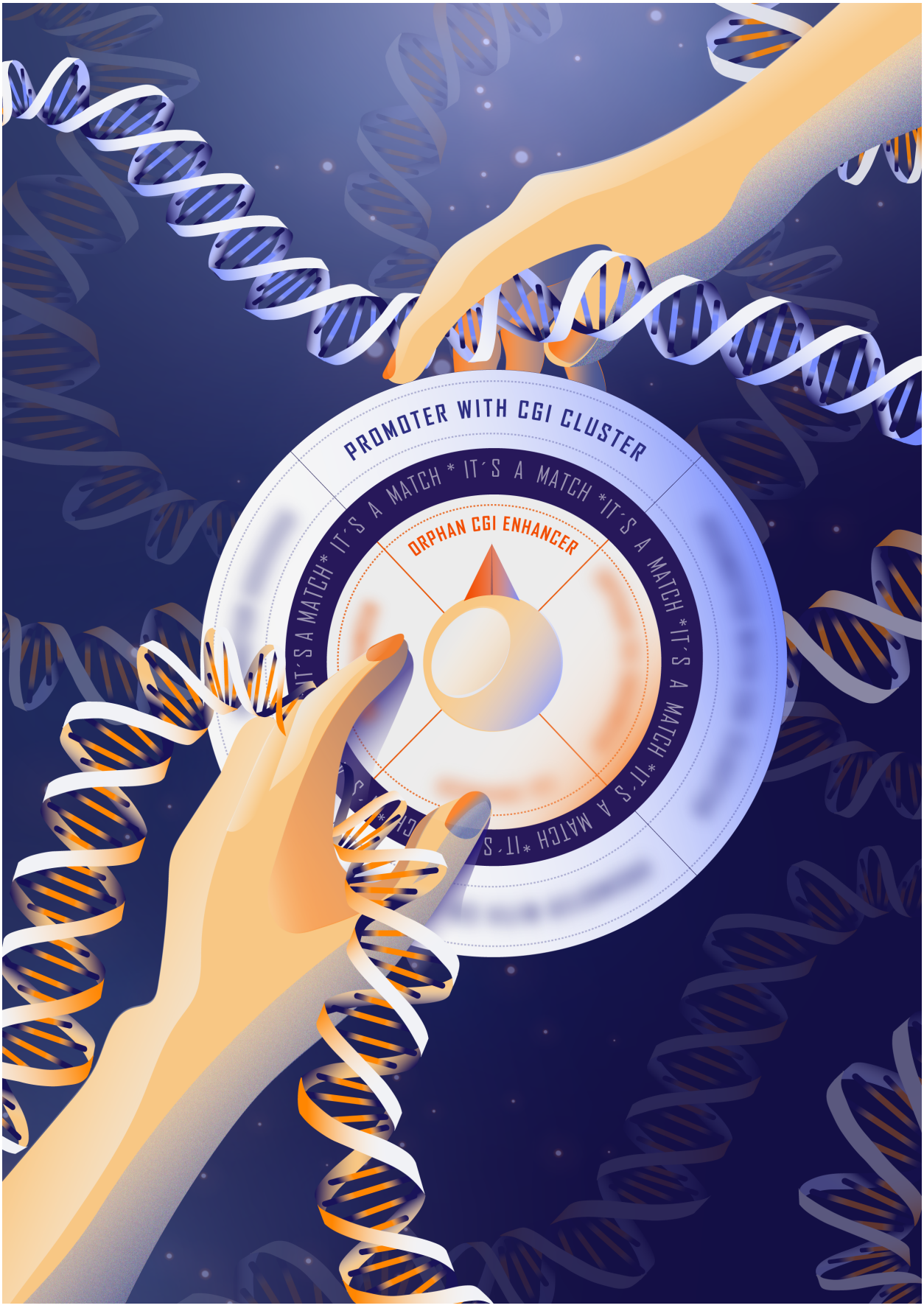

PROMOTER WITH CGI CLUSTER
IT´S A MATCH * IT´S A MATCH * IT´S A MATCH * IT´S A MATCH * IT´S A MATCH *
ORPHAN CGI ENHANCER

Orphan CpG islands dictate the compatibility between poised enhancers and their target genes

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Tomás Pachano
aus Caseros (Argentina)

Köln 2021

The Doctoral Thesis “Orphan CpG islands dictate the compatibility between poised enhancers and their target genes” was performed at the Center for Molecular Medicine of Cologne, an institution belonging to the University of Cologne, from May 2017 to February 2021.

Berichterstatter:

Prof. Dr. Mirka Uhlirova
Dr. Peter Tessarz
Prof. Dr. Siegfried Roth

Tag der mündlichen Prüfung: 17.01.21

A mis viejos

A mis hermanas

A Susi y también a Luis, Nelly y Norberto

Al Rojo campeón de la Sudamericana 2017

Al Mar Cantábrico y Picos de Europa

A Chorizo que no me dejaba escribir en paz

Acknowledgements

This thesis could not have been possible without the direct and indirect contribution of many people in Germany, Spain and Argentina. Here I will change the language of my writing to give a special dedication to some of them:

Álvaro, te agradezco que me hayas guiado en la realización de este proyecto tan audaz. Gracias por la oportunidad, tus enseñanzas y, sobre todo, por contagiarme tu pasión por hacer ciencia.

A las personas del laboratorio que me ayudaron con los experimentos, quiero agradecerlas de corazón. En especial a Victor, que hizo todos los análisis computaciones; a Thais, quien me dio una ayuda enorme durante los últimos experimentos; y a María, que además de ayudarme con experimentos, realizó la hermosa portada de esta Tésis.

Josefina nuestras discusiones científicas en Colonia también son una parte importante de este trabajo. Me siento muy afortunado de haber pasado mis primeros anos académicos con vos.

Mamá, Papá, Julieta y Belén: se nos hizo difícil estar tanto tiempo distanciados. Por lo menos sirvió para esta Tesis. Espero que estén orgullosos.

This work is based on the following publication:

"Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness". Pachano, T; Sánchez-Gaya, V; Ealo, T; Mariner-Faulí, M; G. Asenjo, H, Respuela, P; Cruz-Molina, S; Muñoz-San Martín, M; Haro, E; van Ijcken, W; Landeira, D; Rada-Iglesias, A. *Nature Genetics* 53, 1036–1049 (2021).

Table of Contents

ABSTRACT

The establishment of cell type specific gene expression programs during vertebrate development is largely dependent on distal cis-regulatory sequences known as enhancers. An important question in enhancer biology is to understand how activating regulatory cues from enhancers are communicated specifically to their correct target gene and not others. It is widely assume that insulator elements can regulate enhancer specificity by restricting their search space to regions located within their same topological associated domain (TAD). However, recent evidences have identified examples in which genes do not respond to enhancers placed into their same TAD. Therefore, we hypothesize that genetic and/or biochemical compatibilities must further regulate how promoters respond to enhancers. Here, we uncover a novel role for CpG islands (CGIs) in dictating the compatibility between genes and distal enhancers. By using a CRISPR/Cas9 knock-in strategy, we inserted poised enhancers (PEs) within topological associated domains (TADs) containing genes with different promoter types. Analysis of the resulting cell lines revealed that developmental genes with CGI-rich promoters are particularly responsive to PEs, and that such responsiveness depends on the presence of an orphan CGI in proximity to the PE sequence. We report here that CGIs can amplify the regulatory activity of PEs by conferring them a permissive topological configuration that increases the responsiveness of target genes with CGI-associated promoters. Our results provide major insights into fundamental and unresolved questions of cis-regulatory control during cell differentiation including: (i) the genetic factors controlling the topological features of enhancers; (ii) the regulatory rules dictating the compatibility and responsiveness between genes and enhancers. Moreover, our data will help to understand the pathomechanisms of human congenital diseases involving structural variants.

ZUSAMMENFASSUNG

Die Etablierung zelltypspezifischer Genexpressionsprogramme während der Wirbeltierentwicklung hängt weitgehend von distalen cis-regulatorischen Sequenzen ab, die als Enhancer bekannt sind. Eine wichtige Frage in der Biologie von Enhancern besteht darin, zu verstehen, wie aktivierende regulatorische Hinweise von Enhancern spezifisch an ihr richtiges Zielgen und nicht an andere übermittelt werden. Es wird allgemein angenommen, dass Isolatorelemente die Enhancer-Spezifität regulieren können, indem sie ihren Suchraum auf Regionen beschränken, die sich innerhalb ihrer gleichen topologischen assoziierten Domäne (TAD) befinden. Jüngste Beweise haben jedoch Beispiele identifiziert, in denen Gene nicht auf Enhancer reagieren, die in ihr gleiches TAD eingebracht werden. Daher stellen wir die Hypothese auf, dass genetische und/oder biochemische Kompatibilitäten das Ansprechen des Promotors auf Enhancer weiter regulieren müssen. Hier entdecken wir eine neue Rolle von CpG-Inseln (CGIs) bei der Bestimmung der Kompatibilität zwischen Genen und distalen Enhancern. Mit einer CRISPR/Cas9-Knock-in-Strategie fügten wir balancierte Enhancer (PEs) in topologisch assoziierte Domänen (TADs) ein, die Gene mit unterschiedlichen Promotortypen enthalten. Die Analyse der resultierenden Zelllinien zeigte, dass Entwicklungsgene mit CGI-reichen Promotoren besonders auf PEs ansprechen und dass eine solche Reaktionsfähigkeit von der Anwesenheit eines verwaisten CGI in der Nähe der PE-Sequenz abhängt. Wir berichten hier, dass CGIs die regulatorische Aktivität von PEs verstärken können, indem sie ihnen eine permissive topologische Konfiguration verleihen, die die Reaktionsfähigkeit von Zielgenen mit CGI-assoziierten Promotoren erhöht. Unsere Ergebnisse liefern wichtige Einblicke in grundlegende und ungelöste Fragen der cis-regulatorischen Kontrolle während der Zelldifferenzierung, einschließlich: (i) die Bedeutung von Enhancer-Promotor-Pre-Looping, um einheitliche und präzise Genexpressionsprofile während Entwicklungsübergängen zu etablieren; (ii) die genetischen Faktoren, die die topologischen Merkmale von Enhancern kontrollieren; (iii) die regulatorischen Regeln, die die Kompatibilität und Reaktionsfähigkeit zwischen Genen und Enhancern diktieren. Darüber hinaus werden unsere Daten dazu beitragen, die

ZUSAMMENFASSUNG

Pathomechanismen angeborener Erkrankungen des Menschen mit strukturellen Varianten zu verstehen.

1- INTRODUCTION

1.1 Regulatory elements in development

Development is a complex process involving an enormous amount of molecular reactions to generate functional organisms, yet needs to be very precise and robust to avoid pathogenic phenotypes. For this, the genome utilizes different regulatory elements, which work together to ensure that genes are expressed at specific time, locations and levels. Our current knowledge of the genome identifies three main regulatory elements: (i) promoters, (ii) enhancers and (iii) insulators:

(i) Promoters (further discussed in section 1.2) are cis-regulatory elements which overlap transcription start site (TSS) of coding and non-coding genes (Figure 1.1) (Haberle & Stark, 2018; Lenhard et al., 2012; Shlyueva et al., 2014). They served as a recruitment platform for the transcriptional machinery, which is responsible for a fundamental biological process: the transcription of the DNA sequence into a complementary RNA transcript. These transcripts will then be used as a template to build proteins (protein-coding transcripts) or to regulate other cellular functions (non-coding transcripts).

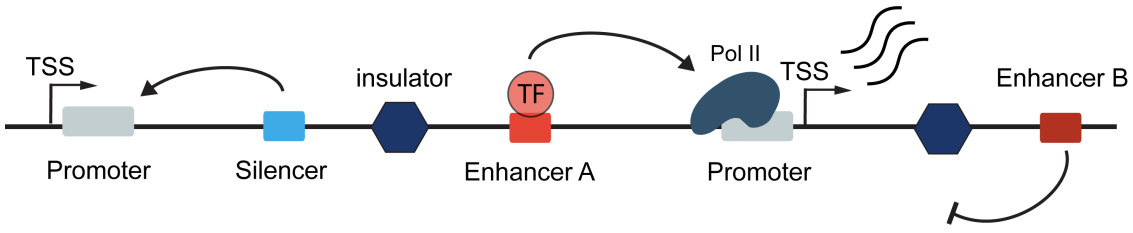


Figure 1.1. The role of regulatory elements in transcription regulation. Promoters are DNA sequences located next to the TSS of a gene, which recruit the RNA polymerase II (Pol II) and general transcription factors (GTFs). Enhancers are cis-regulatory sequences that contain binding sites for TFs that can upregulate the transcription of a target gene. Silencers are cis-regulatory elements that contain

binding sites for proteins that can repress the transcription of a target gene. Insulators are sequences bind by architectural proteins, which can insulate a DNA region from other regulatory elements.

(ii) Enhancers (further discussed in section 1.3) are cis-regulatory elements that contain cognate sequences for DNA-binding proteins, referred as transcription factors (TFs) (Figure 1.1) (Levine, 2010; Muller & Lindmark, 1976; Shlyueva et al., 2014). The union of TFs to enhancers leads to the recruitment of coactivators, which are enzymes that can recruit or activate the transcriptional machinery located at the promoter (e.g. p300, Mediator) (Reiter et al., 2017; Serebreni & Stark, 2021; Spitz & Furlong, 2012).

(iii) Insulators (further discussed in section 1.4) are DNA elements that serve as scaffold for architectural proteins that participate in regulating the 3D configuration of the genome (Figure 1.1) (Bonev & Cavalli, 2016; Rowley & Corces, 2018). One of the main roles of insulators is to partition the genome into topological associated domains (TADs), which reduce the physical distance of co-regulated elements and insulate them from elements with different spatio-temporal activities (Bonev & Cavalli, 2016; Rowley & Corces, 2018).

1.2 Promoters

Promoters recruit the pre-initiation complex (PIC), consisting in RNAP2 and general TFs, and drive transcription initiation at a precise nucleotide position and direction (Haberle & Stark, 2018; Lenhard et al., 2012; Shlyueva et al., 2014). Promoters can initiate transcription on their own but only at low levels (Haberle & Stark, 2018). This basal transcription is usually regulated by other cis-regulatory elements, known as enhancers and silencers, which can further activate gene expression or repress it, respectively (Long et al., 2016; Shlyueva et al., 2014).

1.2.1 Promoter types

On the basis of their genetic composition, transcription pattern initiation and their chromatin marks, metazoan promoters were classified into three main groups (Haberle & Stark, 2018;

Lenhard et al., 2012) (Table 1.2.1): Type I (i.e. promoters of tissue-specific genes), Type II (i.e. promoters of housekeeping genes) and Type III (i.e. promoters of developmental genes). In vertebrates, the main genetic feature that differentiates these promoters is the presence and density of sequences containing concentrated CpG dinucleotides, referred as CpG islands (CGIs). Type I promoters do not contain CGIs and Type II promoters only contain a single and narrow CGI close to the TSS, while Type III promoters usually contain several CGIs, usually extending into the coding sequence of their associated genes.

Table 1.2.1. Properties of vertebrate promoter types.

Type	**Associated gene function**	**Transcription initiation pattern**	**Genetic composition in vertebrates**	**Chromatin marks**
Type I	Tissue-specific	Single-well defined TSS	- Do not contain CGIs. - Contain TATA-box and Inr motifs	- When active, they contain H3K4me3 and H3K27ac - When inactive, they usually do not contain specific chromatin marks
Type II	Housekeeping	Multiple dispersed TSS	- Contain a single CGI. - Do not contain TATA-box motif.	- They contain H3K4me3 and H3K27ac
Type III	Developmental transcription factors	Multiple TSS	- Contain CGI clusters, usually overlapping the coding sequence. - Typically encountered in TADs containing a high number of enhancers.	- When active, they contain H3K4me3 and H3K27ac - When inactive, they are enriched H3K27me3

1.2.2 CpG islands

Among vertebrate genomes, the main distinctive feature to classify promoters is the presence of CGIs (Bajic et al., 2006; Deaton & Bird, 2011; Lenhard et al., 2012). CGIs are genetic elements with high content of hypomethylated CpG dinucleotides. Originally, CGI were identified by using the restriction enzyme HpaII (A. Bird et al., 1985; D. Cooper et al., 1983). After digesting genomic DNA with HpaII, a small portion of the genome was found to resist the enzymatic digestion. Characterization of these sequences revealed that they were rich in guanine and citocine (GC) and with a particularly high CpG content, in contrast with the rest of the genome (A. Bird et al., 1985). Faced with the need of a working definition, Gardiner-Gardner defined CGIs as DNA sequences with a CG% average of ~50%, an observed/expected ratio CpG ratio of over 0.6 and more than 200bp of length (Gardiner-Garden & Frommer, 1987). Then, whole-genome studies were complemented with the identification of individual CGIs in gene loci and it was found that many genes with known hypomethylated promoters were highly enriched in CGIs (McKeon et al., 1982; Stein et al., 1983). This correlation put CGI in the focus of many researchers, who still today try to explain the functionality that CGIs might have in the regulation of their associated genes.

One of the functions of CGIs is to protect DNA from CpG methylation, which, in the context of promoters, might prevent gene silencing (A. P. Bird & Wolffe, 1999). This observation is compatible with the idea that a key function of CGIs it to provide CGI-associated genes with a transcriptional permissive landscape. Several findings support this hypothesis: (i) RNAP2 is bound to inactive CGI-rich promoters (Ferrai et al., 2017; Guenther et al., 2007), (ii) several TF binding sites (TFBS) are over-represented in CGIs (S. A. Jaeger et al., 2010; Landolin et al., 2010), (iii) CGI chromatin is intrinsically more accessible than non-CGI chromatin (Ramirez-Carrozzi et al., 2009), (iv) they are enriched in H3K3me3, a histone mark that favor the recruitment of the transcriptional machinery and protects DNA from methylation (van Ingen et al., 2008; Vermeulen et al., 2007), (v) they are depleted in H3K36me2, a histone mark that

inhibits transcription by recruitment of a DNA methyltransferase (Dhayalan et al., 2010), and (v) they bind chromatin remodellers (Wysocka et al., 2006) and coactivators (Saksouk et al., 2009). The sum-up of these epigenetic features might facilitate the rapid induction of CGI-rich genes during developmental transitions. Interestingly, as mentioned before, most CGI-rich genes are developmental transcription factors (Haberle & Stark, 2018; Lenhard et al., 2012).

Since CGI do not share common and recurrent motifs, how is it possible that all CGI-rich promoters share the similar epigenetic profiles? This can be explain by the fact that many chromatin remodellers contain a zinc finger CxxC (CxxC) domain that specifically recognized non-methylated CpG dinucleotides (Table 1.2.3) (Long, Blackledge, et al., 2013). The CxxC domain-containing protein KDM2B is one of particular importance since it was suggested to be involved in the recruitment of the very well studied polycomb group complexes (PcGs) (Blackledge & Klose, 2021; Farcas et al., 2012; He et al., 2013; Wu et al., 2013). PcGs are sub-classified into two main complexes, Polycomb repressive complex 1 (PRC1) and Polycomb repressive complex 2 (PRC2), which mark inactive CGI-rich promoters with H2AK119ub and H3K27me3, respectively (Blackledge & Klose, 2021). The highly overlapping occurrence of both PRC1 and PRC2 with CGIs is explained in a model where non-canonical PRC1 complexes are recruited to CGI via KDM2B, where they deposit the H2AK119ub mark (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). Subsequently, this histone modification modulates the recruitment and activity of PRC2 complexes through interaction with the JARID2 subunit (Blackledge et al., 2014; Blackledge & Klose, 2021; S. Cooper et al., 2014). Finally, enzymatic activity of PRC2 leads to deposition of H2K27me3 mark, which in turn serve as a docking site for canonical PRC1 binding (Blackledge & Klose, 2021; Zhen et al., 2016). Moreover, PRC2 can also be directly recruited to CGIs via the MTF2 subunit (Perino et al., 2018). Together, this feedback recruitment mechanism between both complexes allow the spreading of PcGs alongside CGI forming the so-called PcG-domains. Although the functional role of PcG-domains is still under debate, it was proposed that they help to protect inactive genes against low-level premature transcription (Akasaka et al., 2001; Blackledge & Klose, 2021).

Table 1.2.3. CxxC domain-containing proteins and their molecular functions

rotein	Molecular function
KDM2A	Removal of H3K26me3 mark(Tsukada et al., 2006)
KDM2B	Recruitment of non-canonical PRC1 and protection from DNA methylation(Boulard et al., 2015; Farcas et al., 2012; He et al., 2013; Wu et al., 2013)
MLL1	Implementation of H3K4me2 and H3K4me3(Hu et al., 2017)
MLL2	Implementation of H3K4me2 and H3K4me3(Hu et al., 2017)
TET1	Demethylation of 5-methylcytosine (5mC) and oxidation of 5mC into 5-hydroximethylcitosine
TET3	Demethylation of 5mC and oxidation of 5mC into into 5-hydroximethylcitosine (restricted expression in ocyte and zygote)
CFP1	Part of SET1 complexes, implicated in the deposition of H3K4me3 at CGIs (Clouaire et al., 2012)

Bird and colleagues made use of the binding specificity of the CxxC domain to biochemically detect CGIs by a method referred as CAP (R. Illingworth et al., 2008). Briefly, they applied fragmented gDNA to a chromatography system with a sepharose matrix linked to a CxxC domain. This matrix can enrich the DNA fragments for unmethylated CpG rich sequences, which can then be interrogated by a microarray. CAP was simplified by Klose and colleagues by replacing the chromatography step by a streptavidin-biotin purification (Blackledge et al., 2012). These techniques allowed the identification of a larger fraction of CGIs than what presumed by computational predictions, including CGIs from from other vertebrate species beyond mammals (Gardiner-Garden & Frommer, 1987). Interestingly, many of these newly discovered CGI were found to be located outside of mouse and human gene promoters (R. S. Illingworth et al., 2010).

These CGIs were termed as orphan CGIs (oCGIs) and, although they correspond to about half of all CGIs, their functional role remains enigmatic (Bell & Vertino, 2017; R. S. Illingworth et al., 2010; Maunakea et al., 2010).

1.2.3 Promoter responsiveness

The genetic diversity of promoters (e.g. CGI content and motif composition) suggest the possibility that each promoter type have a preference to specific enhancers. That is, that there might be genetic rules which regulate compatibility between promoters and enhancers. This was firstly assessed by Butler and Kadonaga in 2001, who generated *Drosophila* lines with random integrations of reporter genes containing a DPE- and TATA-motif promoter (Butler & Kadonaga, 2001). By a selective excision of either one or the other promoter motif, they revealed that some enhancers were able to induced the expression of the reported gene only when they contain a specific promoter type. That is, the responsiveness of the reported genes was influenced by the compatibility between the locus of integration and the promoter type. More than 10 years later, the development of a multiplexed ectopic assay referred as self-transcribing active regulatory region sequencing (STARR-seq) allowed to test the activity of millions of *Drosophila* enhancer candidates towards housekeeping and developmental promoters (Zabidi et al., 2015). Strikingly, the enhancer candidates showed a preference towards one or the other class of promoters. Similarly, a variant of the STARR-seq, referred as self-transcribing active core promoter-sequencing (STAP-seq), determined a wide-range of responsiveness for thousands of candidate core promoters tested against fixed enhancer sequences (Arnold et al., 2016). Moreover, STAP-seq was also used to test how different coactivators induce thousands of different *Drosophila* promoter types (Haberle et al., 2019). In line with the previous results, they found distinct coactivators-promoter compatibilities. Altogether, these results suggest that promoters have specificities towards some activation signals. That is, there are genetic and biochemical compatibility rules that regulate promoter responsiveness to enhancers. However, it

is important to note that in reporter assays enhancers are placed at small distances from reporter genes and in a non-chromatinized environment.

1.3 Enhancers

Enhancers bind tissue-specific TFs that regulate the transcription from genes independently of their relative distance and orientation (Levine, 2010; Long et al., 2016). This enables a gene to be regulated by multiple enhancers with different spatiotemporal activities, which in turn allows a vast amount of cellular transcriptional states with a minimal set of genes (Figure 1.3A). This concept is well exemplified by the *even skipped* (*eve*) enhancers, which are activated in the early *Drosophila* embryo in an alternating stripe manner (Stanojevic et al., 1991). In this example, seven *eve*-expressing stripes are created along the embryo by the activity of five enhancers, being each enhancer responsible for the expression of *eve* in one or two stripes (Figure 1.3B). *Eve* enhancers contain cognate sequences for activating and repressing TFs referred as gap proteins, which form opposing gradients along the embryo anteroposterior axis. Depending on the relative abundance of activating and repressing gap proteins and their corresponding cognate sequences in the enhancers, each enhancer will become active in a specific axis of the embryo (Figure 1.3B).

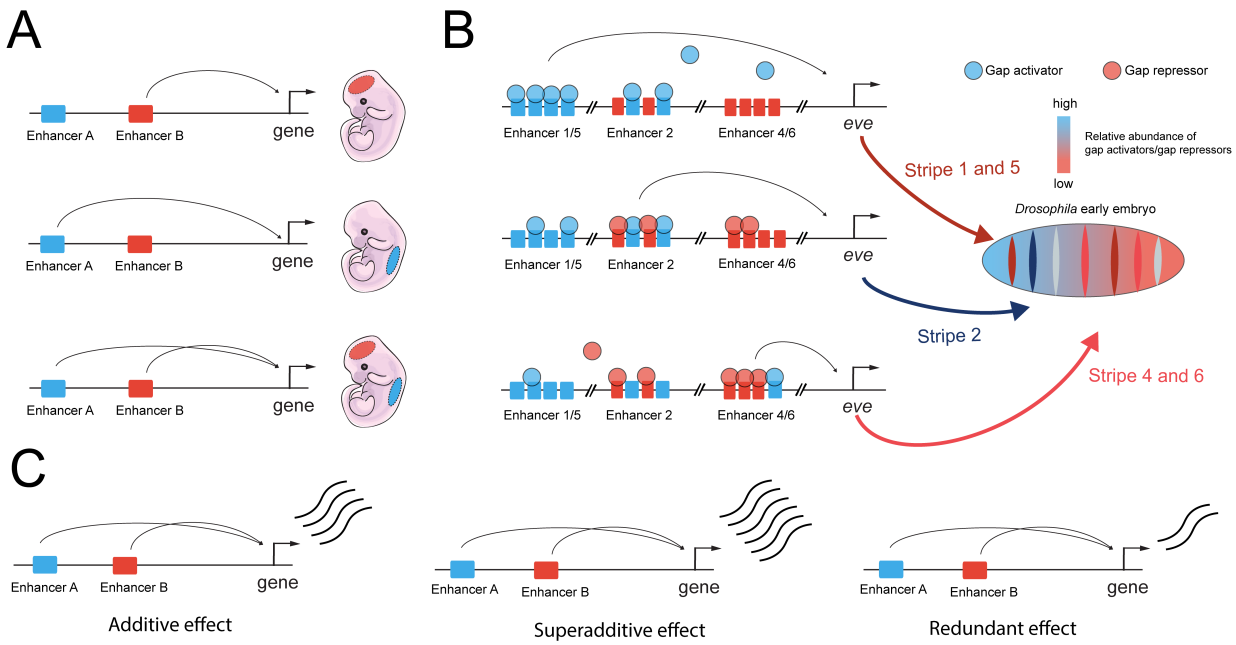


Figure 1.3. Enhancers can function through complex and combinatorial mechanisms of actions. (A) Enhancers are able to confer different spatiotemporal gene expression patterns for the same target gene. **(B)** Expression of *eve* in *Drosophila* early embryo is controlled by the action of different enhancers. Concentration gradients of Gap activator (blue) and Gap repressor proteins (red) determine the specific activation of each enhancer in different parts of the anteroposterior axis of the embryo. **(C)** Enhancers from the same gene can interact in different ways: (left panel) they can function independently where each enhancer module produces a transcriptional output; (center panel) they can interact in a synergistic way to drive more expression than the sum of each enhancer module; or (right panel) they can have a redundant effect on the expression of their target gene.

As more cis-regulatory elements are discovered, it is becoming clear that enhancers usually function in an even more complex combinatorial way (Kvon et al., 2021). As a matter of fact, most genes are regulated by multiple enhancers with often overlapping spatiotemporal activities (Figure 1.3C). The *Drosophila* enhancers that control *knirps* are a classical example illustrating this scenario (Bothma et al., 2015). During late stages of development, the *knirps* enhancers act in an additive manner. That is, the activity of each enhancer independently sum towards *knirps* expression levels. Interestingly, during early stages, both enhancers produce more *knirps* expression levels than what would be predicted from the sum of both. This transcriptional effect

is known as superadditive. This example particularly shows that multiple enhancers can interact in different manners over their target gene depending on the cellular context. It was also shown that many genes, both in mammals and insects, use multiple enhancers with redundant effects (Frankel et al., 2010; Montavon et al., 2011; Osterwalder et al., 2018; Perry et al., 2010, 2011) (Figure 1.3C). As an example, serial deletion of *Hoxd13* enhancers in the developing mouse limb have little effects on *Hoxd13* (Montavon et al., 2011). The lack of transcriptional changes upon deletion of some enhancers (Montavon et al., 2011; Osterwalder et al., 2018) raises the question of what is the functionality of these seemingly dispensable enhancers. One proposed answer is that enhancer redundancy serves as a mechanism for improving transcriptional robustness and precision (Kvon et al., 2008; Lagha et al., 2012). That is, having multiple redundant enhancers might increase the probability of cis-activation, which in laboratory conditions might only be noticeable under physiological or genetic stress (Osterwalder et al., 2018).

1.3.1 Identification of enhancers

Since enhancers can be located far-away from their target genes, identification of enhancer sequences in the genome has been a significant challenge. The first attempts to localize enhancers were made by placing sequential DNA fragments upstream of a reporter gene and testing their ability to activate this gene, an assay referred as “enhancer bashing” (Halfon, 2019; Schaffner, 2015) (Figure 1.3.1A). This principle has been adapted to the next-generation sequencing era and massive parallel reporter assays are now able to test enhancer activities in a huge number of DNA sequences (Arnold et al., 2016; Neumayr et al., 2019).

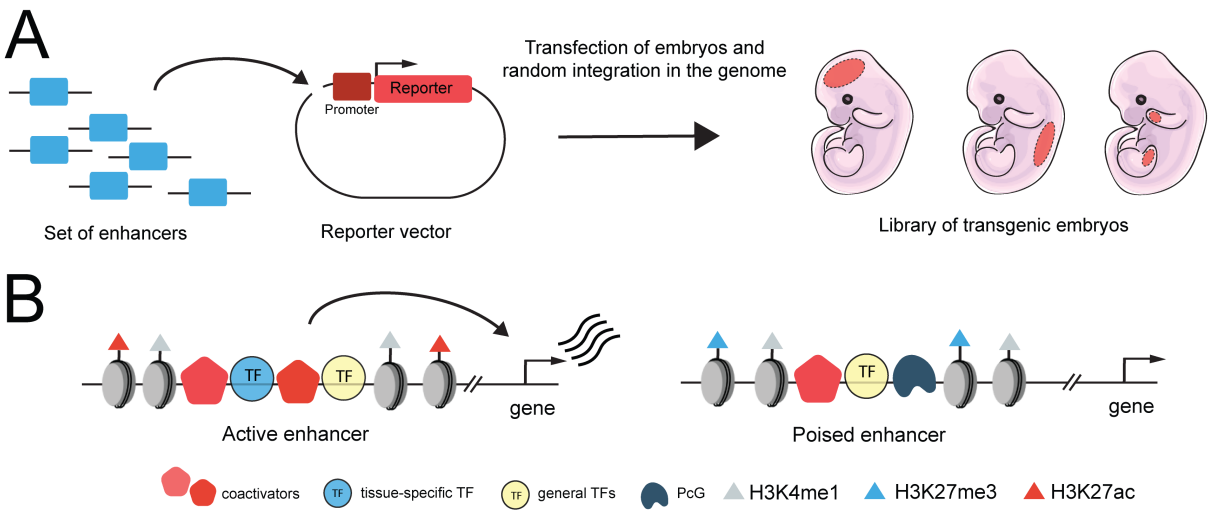


Figure 1.3.1. Enhancers can be identified by different approaches. **A)** A large set of DNA sequences is introduced into a reporter vector and transfected into embryos. The animals can then be screen for enhancer activity by reading the reporter output. **B)** Enhancers can be identified by their chromatin landscape. Active enhancers (left panel) present binding of tissue-specific TFs, coactivators and the histone marks H3K27ac and H3K4me1. Poised enhancers (right panel) present binding of pluripotency TFs (e.g. Oct4, Nanog), coactivators, PcG and the histone marks H3K27me3 and H3K4me1.

Although reporter assays might be useful for identifying new regulatory elements, enhancers are placed in a non-native environment and their context dependent information is neglected (i.e. distance between enhancer and promoter, promoter type and presence of other regulatory elements, chromatin context). However, enhancers sequences can be predicted in their native genomic context by their chromatin signatures (Bonn et al., 2012; Heintzman et al., 2007; Rada-Iglesias et al., 2011; Zentner et al., 2011). Active cis-regulatory elements are typically nucleosome depleted, marked with certain histone modifications and bound by coactivators. Accordingly, transcriptionally active regulatory elements can be identified by ATAC-seq, a method that assess for accessible chromatin sites (Buenrostro et al., 2013), or by ChIP-seq data from histone marks or transcription factors (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Visel et al., 2009). In particular, active enhancer regions are enriched in the monomethylation of

histone 3 lysine 4 (H3K4me1), acetylation of histone 3 lysine 27 (H3K27ac) and binding of p300 (Figure 1.3.1B).

Although enhancers become active only in specific cellular context, some enhancers can also be identified by its chromatin marks before becoming active (Rada-Iglesias et al., 2011; Zentner et al., 2011). Such is the case of Poised enhancers (PEs) (further discussed in section 1.5), which, like active enhancers, display an accessible chromatin, binding of p300 and the H3K4me1 mark (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011) (Figure 1.3.1B). However, PEs are not marked with H3K27ac but with H3K27me3, a histone modification mediated by PcGs (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011).

1.3.2 Mode of action of enhancers

When tissue-specific TFs are expressed, their binding to cognate TF binding sites (TFBS) in enhancer regions result in the recruitment of coactivators (Reiter et al., 2017; Serebreni & Stark, 2021; Shlyueva et al., 2014; Spitz & Furlong, 2012). Coactivators influence the RNAP2 recruitment, initiation or elongation (Shlyueva et al., 2014; Spitz & Furlong, 2012). Therefore, they serve as a functional bridge between the TFs and the transcriptional machinery located at the enhancer and at the promoter, respectively.

Some coactivators mediate post-translational modifications of target proteins, like the acetyltransferase p300. Recruitment of p300 by a TF could result in the acetylation of histone tails, loosen of histones interactions with DNA and consequently favor the binding of other TFs (Lee & Workman, 2007). In addition, it was shown that p300 can directly acetylate the RNAP2 (Schröder et al., 2013), which promotes the conversion of the RNAP2 into an active state that transcribes RNA. Another well studied coactivator recruited by active enhancer elements is the Mediator complex. Mediator can stimulate the phosphorylation of the RNAP2 carboxy-terminal domain (CTD) and, consequently, in the release of the RNAP2 from a transcriptional initiation state towards a productive-elongation state (Kim et al., 1994). Interestingly, depletion of Mediator not only results in the general downregulation genes, but also in the disassembly of nuclear structures containing large clusters of RNAP2 condensates (M. G. Jaeger et al., 2020;

Zamudio et al., 2019). This findings lead to a model that implicates Mediator subunit MED1 with its intrinsically disordered regions in the formation of phase-separated condensates that positively influence transcriptional reactions (Mir et al., 2019).

Altogether, transcriptional regulation is achieved by integrating the input from TF-DNA interactions at the enhancer elements and the transcriptional machinery. Importantly, the transcriptional machinery is positioned not only at the gene promoter but also at the enhancer. Therefore, its activation results in RNAs transcribed from the TSS and from the enhancer, which are referred as the mRNA and enhancer RNA (eRNA), respectively. Whereas mRNAs are translated into proteins, the functional role of eRNA is less clear. There is still a debate whether the production of eRNAs merely reflects transcriptional noise, or whether it has a regulatory function. In support of the latter, the insertion of RNAP2 terminator within the enhancer region that controls the expression of the human growth hormone gene cluster blocks eRNA production and represses the expression of the adjacent gene (Ho et al., 2006). Similarly, using small interfering RNAs to knock-down the eRNA expression of an enhancer that regulates the mouse gene *KLK3* resulted in specific down-regulation of target genes (Hsieh et al., 2014; Melo et al., 2013). Other studies, have also suggested that eRNAs might regulate the RNAP2 binding at target promoters (Mousavi et al., 2013) or enhancer-gene looping (Li et al., 2013; Pnueli et al., 2015). However, these studies are still under debate and were challenged by other studies, which contrary show that some eRNAs do not affect target gene expression (Engreitz et al., 2016; Lloret-Llinares et al., 2018; Paralkar et al., 2016).

eRNA might or might not have a regulatory role towards mRNA production, but both processes seem to be coupled (Andersson et al., 2014; Fitz et al., 2020; Henriques et al., 2018) and independent of the local chromatin activation (Fitz et al., 2020). In support of this, depletion of the elongation factor Spt5 in mouse B cells results in enhancer and gene loss of transcription, but do not affect histone acetylation, chromatin accessibility or Mediator binding (Fitz et al., 2020). Thus, enhancer-gene activation could involve two independent steps: (i) the TF binding to cognate sequences at the enhancer, which bring coactivators and open up the chromatin; and (ii) the transfer of the regulatory information to the promoter, which will activate the transcriptional

machinery at both sites. It is tempting to predict that for the transition from step (i) to (ii) a physical communication between enhancers and promoters must happen. However, recent studies using super-resolution microscopy revealed a lack of correlation between transcription and enhancer-gene contacts (Alexander et al., 2019; Benabdallah et al., 2019; Mateo et al., 2019). In light of these findings, new models for contact-mediated cis-regulation of gene expression were proposed (Xiao et al., 2020; Zuin et al., 2021). These models postulate that a small increase in enhancer-promoter contact could produce a larger change in gene expression by an hysteretic reaction from the promoter. That is, regulatory steps will be occurring at a specific rate when the enhancer is in close contact with the promoter. After an accumulation of *n* contacts in a short time, the occurring rate will pass a threshold barrier and the transcriptional machinery will switch to an active state, where it can maintain transcription for some time without the contact of the enhancer.

1.4 Genome organization

The conclusion of the previous sections is that transcriptional regulation is achieved by the transfer of activating cues from the enhancer to the target gene promoter. But how do enahncers specifically communicate with their target genes and not others? A key insight into this question came from studies using chromatin conformation capture (3C) technologies, which enable the detection of physically interacting loci in a cell population (Dekker et al., 2002; Denker & De Laat, 2016; Schmitt et al., 2016). The genome-wide profiles of these interactions revealed the existence of spatially restricted chromatin domains, referred as topological associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012). TADs are characterized by high intra-domain interactions and low interactions with sequences in their neighbouring domains (Figure 1.4.1A) (Dixon et al., 2012; Nora et al., 2012). The boundaries sequences that delimit TADs are determined by the probability of the sequence to act as a chromatin insulator in a population level and are enriched in cognates motifs for the insulator protein CCCTC-binding factor (CTCF). So far, the best mechanistic explanation of TAD formation is the loop-extrusion model

(Fudenberg et al., 2016). This model postulates that the architectural complex, Cohesin, is loaded into chromatin and progressively extrudes a loop, until it encounters a chromatin-bound CTCF anchor loop formed by convergent CTCF binding sites (CBS) (Fudenberg et al., 2016; Merkenschlager & Nora, 2016; Sanborn et al., 2015) (Figure 1.4.1A,B). Consequently, the looping extrusion will bring together loci from the same TAD, but rarely loci from adjacent TADs.

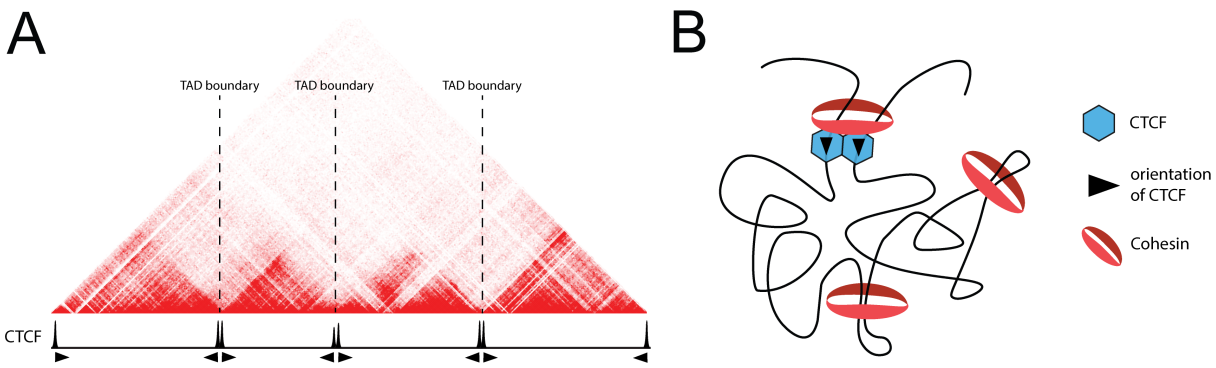


Figure 1.4.1. TAD formation by loop extrusion. A) Hi-C profiles from mESC (Bonev et al., 2017), showing a representative locus containing 4 TADs. CTCF ChIP signals(Pope et al., 2014) are pictured as peaks with black arrows below representing the orientation of the CTCF. B) Diagram representing the structure of the TAD according to the loop extrusion model: Cohesin rings form loops while "scanning" the chromatin, but are stalled by convergent CTCF anchor loops.

TAD boundaries have been postulated to demarcate the enhancer search space to sequences that are located inside the domain (Dowen et al., 2014; Hanssen et al., 2017). This scenario is supported by the studies of genomes with structural variations involving TAD boundaries (i.e. deletions, duplications, inversions or translocations) (Lupiáñez et al., 2016; Smith & Shilatifard, 2014; Spielmann et al., 2018). Deletion of TAD boundaries can result in the extension of TADs, which can lead to the new enhancer-gene interactions and consequently to ectopic gene

activation (i.e. enhancer adoption or hijacking) (Franke et al., 2020; Gröschel et al., 2014; Hanssen et al., 2017; Lupiáñez et al., 2015) (Figure 1.5A). One case is the removal of a CTCF binding site (CBS) at the α-globin locus, which drives the fusion between two neighbouring TADs, ectopic interactions of the α-globin enhancers and upregulation of genes present in the extended TAD (Hanssen et al., 2017). On the other hand, inversions can "shuffle" TADs, which can lead to gaining of gene expression by enhancer adoption and/or loss of gene expression by enhancer disconnection (Laugsch et al., 2019; Mehrjouy et al., 2018; Redin et al., 2017) (Figure 1.4.2A). For example, an inversion found in patients with abnormal limb development relocates the enhancers from *EPHA4* into the TAD of *WNT6* (Lupiáñez et al., 2015). *C*onsequently, this leads to the ectopic activation of WNT6 in a *EPHA4*-like pattern and loss of *EPHA4* expression.

The above mentioned examples suggest that TAD boundaries prevent enhancers from activating non-target genes. Therefore, if TAD boundaries are disrupted there should be a drastic genome-wide effect in gene expression (by a massive enhancer rewiring). However, when CTCF or Cohesin are depleted, only a minor set of genes suffer changes in their expression levels (Nora et al., 2017; Rao et al., 2017b). The same prediction is also challenged by experiments using highly rearranged chromosomes in *Drosophila* (i.e. chromosomes with large and small inversion, duplications and deletions), where only a few genes display significant expression changes (Ghavi-Helm et al., 2019). Consistent with these findings, we have also found a case where genes do not respond to enhancers located in their same TAD after an inversion (Laugsch et al., 2019). This example correspond to an inversion found in a patient with Branchio-oculo-facil syndrom (BOFS) which places the *TFAP2A* gene into a novel TAD. Because of a physical disconnection from its cognate enhancers, *TFAP2* is not properly induced in neural crest cells (Figure 1.4.2B). Interestingly, this inversion also places novel genes in the same "shuffled" TAD with *TFAP2A* neural crest enhancers. However, none of these genes is responsive to *TFAP2A* enhancers (Figure 1.4.2B).

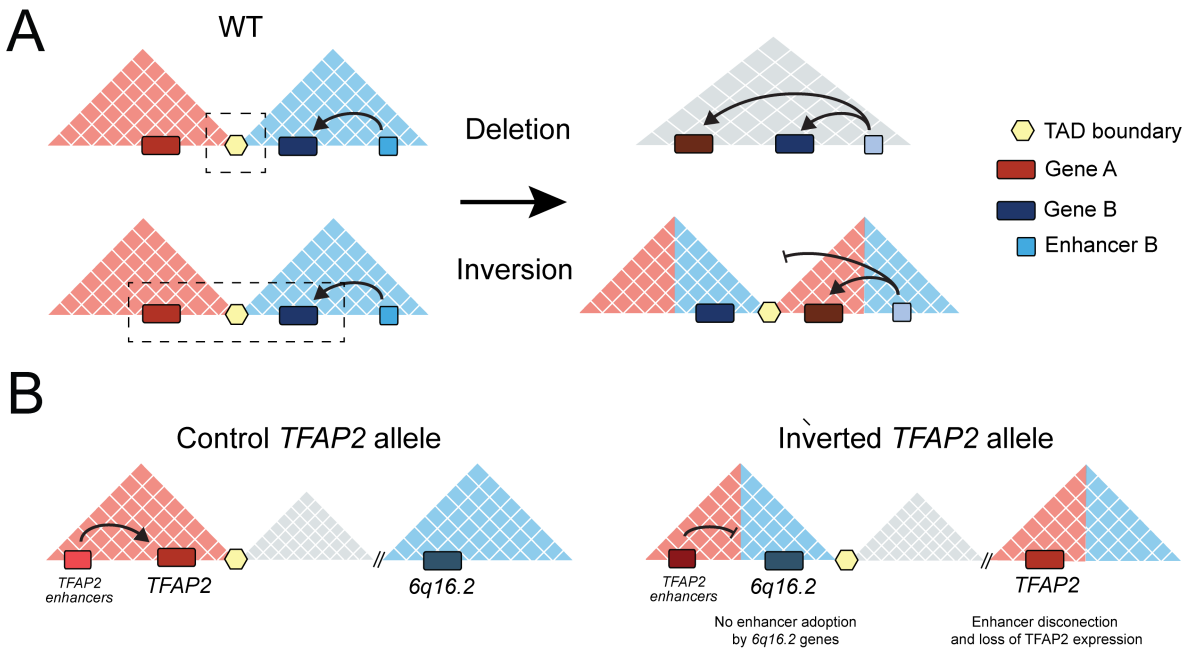


Figure 1.4.2. Structural variants and enhancer rewiring. A) Structural variants spanning TAD boundares can induce ectopic gene expression by enhancer adoption and/or gene silencing by enhancer disconnection. In the upper panel scheme, a deletion compromising a TAD boundary located between the TADs of Gene A and Gene B can induce the fusion of both TADs. Consequently, Enhancer B, which in WT conditions specifically regulates Gene B, is now able to contact and drive ectopic expression of Gene A. In the lower panel scheme, an inter-TAD inversion compromising Gene A and Gene B can cause enhancer adoption of Gene A. At the same time, the inversion disconnects Gene B from Enhancer B and this causes its silencing. B) Example of a structural variant including a TAD boundary which does not result in enhancer adoption (Laugsch et al., 2019). The inversion places one of the *TFAP2A* alleles into a novel TAD and impairs its normal expression in neural crest cells due to the physical disconnection from its enhancers. Moreover, this inversion also places novel genes originally found within the 6q16.2 locus in proximity of the *TFAP2A* neural crest enhancers within a shuffled TAD. Surprisingly, none of the 6q16.2 genes is responsive to the *TFAP2A* neural crest enhancers (*i.e.* no enhancer adoption in neural crest cells).

Altogether, these findings suggest that placing enhancers and genes in the same TAD is not sufficient for generating productive transcription. TAD boundaries might only play a partial role in regulating enhancer specificity, presumably by restricting enhancer search space, but other mechanisms also have to be taken into consideration. In support of this idea, enhancer reporter

assays in *Drosophila* revealed that core promoter from developmental and housekeeping genes display a differential responsiveness to enhancers (Zabidi et al., 2015). Thus, genetic elements found in promoters (and/or enhancers) might contribute to control enhancer-gene compatibility. Deciphering these rules is fundamental to better understand the pathogenic potential of structural variants.

1.5 Poised Enhancers

As previously mentioned, PEs are inactive cis-regulatory elements that are bookmarked with specific chromatin signatures, including H3K27me3, DNA hypomethylation and binding of p300 (Cruz-Molina et al., 2017; Rada-Iglesias et al., 2011). PEs were first described in human and mouse embryonic stem cells (hESC and mESC, respectively) as a small set of highly conserved enhancers associated to the genes which are inactive in ESC but that become induced in the anterior neural linage (Rada-Iglesias et al., 2011; Zentner et al., 2011). A recent study from our lab further demonstrate that PEs are also present in vivo and their features are commonly found in other vertebrates (Crispatzu et al., 2021). The functional relevance of PEs was assessed by loss of function experiments, both in vitro and in vivo (Crispatzu et al., 2021; Cruz-Molina et al., 2017). Deletion of PEs in mESC did not result in transcriptional changes of their target genes (Cruz-Molina et al., 2017). However, when these cell lines where differentiated into anterior neural progenitor cells (AntNPC), PEs target genes failed to get induced (Cruz-Molina et al., 2017). Similar results where shown when some PEs were deleted in mouse and chicken embryos (Crispatzu et al., 2021). Together, these experiments demonstrate that PEs are essential regulatory elements for the proper establishment of the anterior neural program.

PEs most distinctive feature is that they are bound by PcG complexes (i.e. PRC1 and PRC2) and enriched in its associated histone marks (i.e. H2AK119ub and H3K27me3). Since these complexes are commonly associated to transcriptional repression (Blackledge & Klose, 2021), one could expect that the role of PcG in PEs is to maintain these elements inactive. However,

disruption of PRC2 by deletion of the EED subunit did not result in activation of PEs neither their target genes in mESC (Cruz-Molina et al., 2017). Contrary, the induction of these genes was severely hampered after differentiation of EED$^{-/-}$ mESC into AntNPC. Thus, PcG is not necessary to maintain PEs or their target genes inactive in mESC but it might rather act as a facilitator of the induction of these genes (as seen in other contexts (Cohen et al., 2018, 2019; Creppe et al., 2014; Frangini et al., 2013; Gao et al., 2014; Kondo et al., 2014; Loubiere et al., 2020)). The facilitator role of PcG in the context of PEs function might be explained by the topological role of PcG (Pachano et al., 2019). Namely, the disruption of PRC2 or PRC1 impairs the physical association that PEs and their target genes have before activation, referred to as prelooping (Crispatzu et al., 2021; Cruz-Molina et al., 2017). Therefore, PcG-PcG contacts between PEs and their target genes (whose promoters are also enriched in PcG binding) bring together both elements in close proximity and this configuration might facilitate the cis-activation capacity of the enhancer (Figure 1.5). However, it is important to note that this model is based on global loss of PcG, which can elicit genome-wide transcriptional and topological changes and consequently indirectly alter PEs loci.

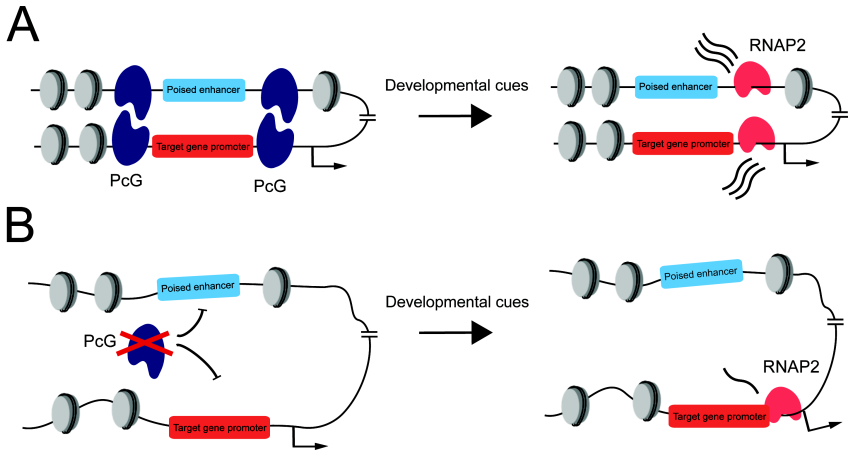


Figure 1.5. PcG as topological facilitator of PE cis-activation. A) In ESC, PcG is present at PEs and their target gene promoters. This probably mediates the physical interaction between both elements. Upon differentiation of ESC into AntNPC, this permissive topology might favour the cis-activation capacity of the PE to activate the physically proximal promoter. **B)** When PcG are disrupted, the prelooping between PE and their target genes is lost. Consequently, PE target gene induction is reduced.

2- AIM

The aim of this project is to dissect the regulatory logic of poised enhancers in order to understand how these elements can specifically communicate with their target genes. Namely, we want to answer: (i) how different types of promoters respond to PEs; (ii) how the different genetic elements from PEs influence the cis-activation capacity, (ii) the local activation and (iii) the topological properties of the PEs. This Thesis provide major insights into how enhancers specifically control the induction of their target genes, which we anticipate will help to better understand pathomechanisms of human structural variants.

3- MATERIAL AND METHODS

3.1 Equipment

Table 3.1. Equipment

Equipment	Company	Catalog number
Bacteria Incubator	Infors HT Ecotron	s00120638
Cell culture Incubator	Sanyo	8070263
Inverted microscope	Leica DMILED	376977
Fluorescent microscope	Olympus IX2-UCB	9A01123
Electrophoretic chamber	Biorad	1704502 and 1704406
Electrophoretic power supply	Biorad	041BR110323
Cell culture centrifuge	Hermle	311001101
Centrifuge for falcon tubes	Hermle	31130026
Centrifuge for Eppendorf tubes	Hermle	61150064
Centrifuge for bacteria culture	Labscience	7.601.314.101
Tissue Culture hood	Kojair	22198
EpiShear Probe Sonicator	Active motif	53051
q-PCR LightCycler 480 II	Roche	5662
c1000 Touch PCR cycler	Biorad	ct024292
Spectrophotometer Nanodrop	Thermo Scientific	F673
Automated cell counter	Biorad	508BR05586
Thermo Block	Ditabis	980052301
ChemiDoc MP	Biorad	731br01726

3.2 Chemicals and reagents

3.2.1 Chemicals

Table 3.2.1. Chemicals

Reagent	Company	Catalog number
2-Propanol	Roth	9866.5
25:24:1 Phenol- Chlorophorm Isoamyl Alcohol	Sigma Aldrich	P2069
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	HN78.2
Adenosine triphosphate (ATP)	Sigma Aldrich	A2383
Agarose	Life Technologies	16500100
Bovine Serum Albumina (BSA)	Roth	3737.2
Chloroform	Sigma Aldrich	366919
DAPI	Sigma Aldrich	D9542-5MG
Dithiothreitol (DTT)	Roth	6908.3
Ethanol	Roth	5054.3
ethylendiaminetetraacetic acid (EDTA)	Roth	8043.1
ethylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid) EGTA	Roth	3054.2
Formaldehide solution 37%	Sigma Aldrich	252549
Glycerin	Sigma Aldrich	68898
Gycerol	Roth	3783.2
Hydrochloric acid (HCl)	Roth	281.1
Litium Chloride (LiCl)	Roth	3739.2
Magnesium Chloride (MgCl2)	Roth	KK36.2
Methanol	Sigma Aldrich	494437

Mounting medium	Southern Biotech	0100-01
N-Lauroylsarcosine	Sigma Aldrich	61743
Na-Deoxycholate	Sigma Aldrich	D678
NP-40	Sigma Aldrich	I3021
Phosphate buffered saline (PBS)	Sigma Aldrich	D8537
Potasium Chloride (KCl)	Roth	HN02.3
Sodium Acetate (C2H3NaO2)	Roth	6773.2
Sodium Chloride (NaCl)	Roth	3957.2
Sodium dodecyl sulfate (SDS)	Roth	183.3
Triton X-100	Roth	3051.4
Trizma Base	Sigma Aldrich	T1503
Tween-20	Roth	9127.2

3.2.2 Buffers and solutions

Table 3.2.2. Buffers

Solution	**Composition**
Blocking Solution	PBS (1x) 0.5% BSA (w/v)
Lysis Buffer 1 (LB1)	50 mM Hepes(pH 7,5) 140 mM NaCl 1 mM EDTA 10% glycerol 0.5% NP-40 0.25% TX-100 dH20
Lysis Buffer 2 (LB2)	10 mM Tris pH 8 200 mM NaCl 1 mM EDTA 0.5 mM EGTA dH20

Lysis Buffer 3 (LB3)	10 mM Tris pH 8 100 mM NaCl 1 mM EDTA 0.5 mM EGTA 0.1% Na-Deoxycholate 0.5% N-lauroylsarcosine dH20
RIPA Wash Buffer	50 mM Hepes pH 7,5 500 mM LiCl 1 mM EDTA 1% NP-40 0.7% Na-Deoxycholate dH2O
Elution Buffer	50 mM Tris pH 8 10 mM EDTA 1% SDS dH2O
PBST washing buffer	PBS 1% Tween
Permeabilizing solution IF	0.3% Triton X-100 dH2O
4C Lysis Buffer	50 mM Tris–HCl pH 7.5 150 mM NaCl 5 mM EDTA 0.5% NP-40 1% TX-100 1X protease inhibitors dH2O
4C Ligation Buffer	50 mM Tris-HCl pH 7.6 10 mM MgCl2 1 mM ATP 1mM DTT dH2O
TE Buffer	10 mM Tris pH 8.0 1 mM EDTA dH2O
TAE buffer (1X)	0mM Tris pH 8.6 20mM Acetate 1mM EDTA dH2O

3.2.3 Kits

Table 3.2.3. Kits

Reagent	Company	Catalog number
Innuprep RNA mini kit	Analytic Jena	845-KS-2040250
EZ DNA Methylation-Direct Kit	Zymo Research	D5020
PCR purification column	QIAgen	28104
ProtoScript II First Strand cDNA Synthesis Kit	NEB	E6560L
Nucleospin Plasmid MiniPrep, 250 rxn	Macherey-Nagel	740588-250

3.2.4 Cell culture reagents

Table 3.2.4.1. Cell culture reagents

Reagent	Company	Catalog number
Tryple Express	Thermo Fisher	25200072
Xav939	Sigma Aldrich	x3004
Puromicine	Thermo Fisher	A2856,0025
Optimem medium	Thermo Fisher	51985026
Fugene HD Transfection reagent	Promega	E2313

Table 3.2.4.2. Cell culture medium

Medium	Component	Company	Catalog
Serum + LIF	15% Fetal Bovine Serum	Life Technologies	16141-061
	1% Antimycotic/Antibiotic	Sigma-Aldrich	A5955-100ML

	0.02% Beta-mercaptoetanol 55 mM	Life Technologies	21985-023
	1 %Glutamax	Life Technologies	35050-038
	1 % MEM NEAA	Life Technologies	11140-035
	Knock-out DMEM	Life Technologies	10829-018
	LIF	N/A	N/A
Freezing medium	40% Serum+LIF medium		
	40% FBS	Life Technologies	16141-061
	10% DMSO	Sigma Aldrich	D2650
N2B27	50% DMEM/F-12	Invitrogen	21041-025
	N2 suplement	Invitrogen	17502-048
	1% penicillinstreptomycin	Invitrogen	15070
	2 mM l-glutamine	Invitrogen	25030-081
	50% Neurobasal medium	Invitrogen	12348-017
	B27 suplement + vitamin A	Invitrogen	12587-010
	0.1 mM Beta-mercaptoethanol	Sigma Aldrich	M6250

3.2.5 Molecular Biology reagents

Table 3.2.5.1. Molecular Biology reagents

Reagent	**Company**	**Catalog number**
Proteinase K	Sigma Aldrich	P2308
RNAse A	Peqlab	12-RA-03
Turbo DNAse	Thermo Fischer	AM1907M
Proteinase Inhibitors	Roche	5892953001
Glycogen	Peqlab	37-1810

QuickExtract DNA Extraction Solution	Biozym	QE09050
dNTPS	Promega	U1515
SYBR® Safe DNA gel stain	Invitrogen	S33102
Dynabeads Protein G	Thermo Fisher	1004D
GeneRuler 1 kb Plus DNA Ladder	Peqlab	25-2240

Table 3.2.5.2. Enzymes

Enzyme	**Company**	**Catalog number**
NlaIII	NEB	R0125L
DpnII	NEB	R0543M
SpeI	NEB	R0133L
MluI	NEB	R0198L
ClaI	NEB	R0197S
NotI-HF	NEB	R3189S
NdeI	NEB	R0111S
T4-Ligase	Invitrogen	15224-041
T4-Ligase	NEB	M0202M
Expand long template PCR system	Roche	11681842001
ORA qPCR Green ROX	HighQu	QPD0150
GoTaq® G2 Hot Start Polymerase	Promega	M7406
AccuPrime™ GC-Rich DNA Polymerase	Thermo Fisher	12337016
TaKaRa EpiTaq HS	Takara	R110A
KAPA HiFi HotStart	Kapa Biosystems	KK2601

3.2.6 Bacteria culture reagents

All vectors were cloned using Top10 E. Coli bacteria strain (gift from Kuran Lab).

Table 3.2.6. Vectors

Vector	Source	Identifier
pX330-hCas9-longchimeric-grna-g2p	Kurian Lab	N/A
AttB-GFP	Addgene	65521
pGEM-T	Promega	A1360

3.3 Cell culture methods

Cell culture was performed under sterile conditions using a laminar flow cell culture hood, which was disinfected after every use with UV light. The solutions used during cell culture were previously autoclaved and the medium used was supplemented with antibiotics and antimitotics. Cell were kept in an incubator at 37°C and 5% CO_2.

3.3.1 Cell lines

For all cell culture methods, WT (E14) mESC were used (gift from Wysocka Lab). In addition, several genetic deletions, insertions and inversions were made using CRISPR-Cas9 (Table 3.3.1).

Table 3.3.1. mESC lines

Cell lines	Identifier
PE Sox1 (+35)$^{-/-}$	Cruz-Molina et al., 2017
PE Sox1 (+35)CGI$^{-/-}$	N/A
PE Sox1 (+35)TFBS insertion in Gata6-TAD	N/A
PE Sox1 (+35)CGI insertion in Gata6-TAD	N/A

PE Sox1 (+35)TFBS+CGI insertion in Gata6-TAD	N/A
PE Sox1 (+35)TFBS insertion in FoxA2-TAD	N/A
PE Sox1 (+35)CGI insertion in FoxA2TAD	N/A
PE Sox1 (+35)TFBS+CGI insertion in FoxA2-TAD	N/A
PE Wnt8b (+21)TFBS insertion in Gata6-TAD	N/A
PE Wnt8b (+21)CGI insertion in Gata6-TAD	N/A
PE Wnt8b (+21)TFBS+CGI insertion in Gata6-TAD	N/A
PE Sox1(+35)TFBS insertion in Gria1-TAD	N/A
PE Sox1 (+35)CGI insertion in Gria1-TAD	N/A
PE Sox1 (+35)TFBS+CGI insertion in Gria1-TAD	N/A
artificial CGI insertion in Gata6-TAD	N/A
artificial CGI+PE Sox1 (+35)TFBS insertion in Gata6-TAD	N/A
PE Sox1(+35)TFBS insertion in Gria1-TAD + pCGI in Gria1-TSS	N/A
PE Sox1(+35)TFBS+CGI insertion in Gria1-TAD + pCGI in Gria1-TSS	N/A
PE Sox1 (+35)TFBS insertion in Gria1-TSS	N/A
PE Sox1 (+35)TFBS+CGI insertion in Gria1-TSS	N/A
PE Sox1 (+35)TFBS insertion in Sox7/Rp1l1-TAD	N/A
PE Sox1 (+35)TFBS+CGI insertion in Sox7/Rp1l1-TAD	N/A
Deletion of Six3/Six2 TAD boundary	N/A
Inversion of Six3 and Six2 genes	N/A
Inversion of Lmx1a and PE Lmx1a (+113)	N/A

3.3.2 Culture of mESC

All mESC (WT and engineered cell lines) were grown on gelatin-coated tissue-culture plates using Serum + LIF medium (Table 3.6).

3.3.3 Differentiation of mESC into AntNPC

Differentiation of mESC in AntNPC was done following a previously described protocol with minor modifications (Gouti et al., 2014). mESC were plated at a density of 6000 cell/cm2 on plates previously coated with Geltrex (Life Technologies) and grown for two days in N2B27 medium supplemented with 10ng/ml of b-Ffg (Table 3.6). Next, cells were grown for one day with 10ng/ml of b-Ffg + 5µM Xav939 and then were grown for two more days only with 5µM Xav939.

3.4 Genetic engineering of mESC using CRISPR-Cas9

Genetic engineering of mESC was performed using CRISPR guide RNAs (gRNAs) and the Cas9 nuclease. Genetic deletions or inversions were performed using two gRNAs flanking the locus of interest. Genetic insertions were performed using one gRNAs targeted to the region of integration and a DNA template with the sequence of interest flanked by homology arms.

3.4.1 Design and annealing of gRNAs

Design of gRNAs was performed with Benchling CRISPR toolkit (www.benchling.com) and only gRNAs that displayed on-target scores >60 and off-target scores <1 were used. Selected gRNAs were synthesized (https://eu.idtdna.com/) as two complementary oligos with the BbsI restriction site overhang at both ends (Table 3.4.1). For the annealing, 1 µl of 100 µM of each oligo were resuspended in 8 µl of water and incubated at 95°C for 5 minutes and subsequent cooling to 25°C at a rate of 5°C per minute.

Table 3.4.1.gRNAs

Name	Oligo Fwd (5´-3´)	Oligo Rev (5´-3´)	gRNA	Coordinates (mm10)

PE-Sox1_deletion left	caccGGCCGGCCTCC TTGGCCCCA	aaacTGGGGCCAAGG AGGCCGGCC	GGCCGGCCTCCTTG GCCCCA	chr8:12,430,502-12,430,521
PE-Sox1_deletion right	caccTGGCCCCCGCA AGAACCTAG	aaacCTAGGTTCTTGC GGGGGCCA	TGGCCCCCGCAAGA ACCTAG	chr8:12,431,547-12,431,566
Gata6-TAD_insertion	caccCTCATAACTCA CGACGCTCC	aaacGGAGCGTCGTG AGTTATGAG	CTCATAACTCACGA CGCTCC	chr18:11,148,342-11,148,361
FoxA2-TAD_insertion	caccGCTCTATGATA AGATCACTG	aaacCAGTGATCTTAT CATAGAGC	GCTCTATGATAAGA TCACTG	chr2:147,926,131-147,926,150
Gria1-TAD_insertion	caccCTAGGCATCAG AACCCCATG	aaacCATGGGGTTCT GATGCCTAG	CTAGGCATCAGAAC CCCATG	chr11:56,909,884-56,909,903
Gria1-TSS_insertion	caccCCGTGCCTCTCC TAAAGATG	aaacCATCTTTAGGA GAGGCACGG	CCGTGCCTCTCCTA AAGATG	chr11:57,010,989-57,011,008
Sox7-TAD_insertion	caccACTTCCAGGAG ACACAACCG	aaacCGGTTGTGTCTC CTGGAAGT	ACTTCCAGGAGACA CAACCG	chr14:63,968,221-63,968,240
Six3/Six2 boundary deletion_left	caccCCCTGTAGTCA CTTGCCCTT	aaacAAGGGCAAGTG ACTACAGGG	CCCTGTAGTCACTT GCCCTT	chr17:85,629,769-85,629,788
Six3/Six2 boundary deletion_right	caccCGTAGCATCAA CAAAGGCCA	aaacTGGCCTTTGTTG ATGCTACG	CGTAGCATCAACAA AGGCCA	chr17:85,666,678-85,666,697
Six3/Six2 boundary inversion_left	caccCGAAGTGTATC GTTTTTTAT	aaacATAAAAAACGA TACACTTCG	CGAAGTGTATCGTT TTTTAT	chr17:85,587,234-85,587,253
Six3/Six2 boundary inversion_right	caccCAGGATGATAT CTGCCGTGC	aaacGCACGGCAGAT ATCATCCTG	CAGGATGATATCTG CCGTGC	chr17:85,697,725-85,697,744
Lmx1a inversion_left	caccGTGTTCAGGTA TAGCCTGAG	aaacCTCAGGCTATA CCTGAACAC	GTGTTCAGGTATAG CCTGAG	chr1:167,597,604-167,597,623
Lmx1a inversion_right	caccACAACACACCA CTATGCCTG	aaacCAGGCATAGTG GTGTGTTGT	ACAACACACCACTA TGCCTG	chr1:167,859,246-167,859,265

3.4.2 Generation of gRNA-CRISPR Cas9 expressing vector

Annealed gRNAs were diluted 1:200 and ligated using T4 ligase for 30 minutes with 50 ng of pX330-hCas9-long-chimeric-grna-g2p vector (Table 3.9), previously digested with BbsI restriction enzyme. Then, 30 µl of chemically competent *E. coli* were thaw on ice and mixed with 2.5 µl of the ligation reaction. The bacteria was then incubated for 1 minute at 37°C and

another minute on ice. Next, 1 ml of LB medium was added to the bacteria, incubated for 1 hour at 37°C with shaking. Subsequently, 300 µl of bacteria was plated in LB-ampiciline plates and incubated overnight at 37°C. The next day, a bacteria colony was picked with a micropipette tip, transferred into a tube with 4 ml of LB-ampiciline medium and grown overnight at 37°C. The vector was purified from the bacteria using a commercial kit (Table 3.4) and validated by Sanger sequencing.

3.4.3 Generation of donor template for genetic insertions

The donor template was generated by the PCR amplification of the donor-cassette vector. The later was generated by ligating: (i) 300 bp homology arms flanking the recognition site of the gRNA (methods 3.5.8.2); (ii) the cassette of interest (methods 3.5.8.2); and (iii) a previously modified attB-GFP vector (Table 3.9) (Figure 3.4.3). The PCR amplification of the donor vector was performed using a forward primer located at the 5` end of the left homology arm and a reverse primer located at the 3´ end of the right homology arm (methods 3.5.8.3). The resulting reaction was purified using a commercial kit (Table 3.4).

Figure 3.4.3. Diagram of the donor-cassette vector. Left homology arm (LHA) was amplified by PCR of WT (E14) and inserted into the backbone vector with NdeI and SpeI restriction enzymes. Right homology arm (RHA) was amplified by PCR of WT (E14) and inserted into the backbone vector with MluI and NotI restriction enzymes. The sequence of interest was amplified by PCR of WT (E14) and inserted between both homology arms using the SpeI and MluI restriction enzymes.

3.4.4 Generation of mESC clones with genetic deletions, inversions or integrations

For generating genetic deletions and inversions, mESC were transfected with 250 ng of each gRNA-CRISPR Cas9 expression vector. For genetic integrations, mESCs were transfected with 250ng of a single gRNA-CRISPR Cas9 expression vector plus 500 ng of the donor template. All transfections were performed with Lipofectamine 3000 following the kit manufacturer protocol. The day after transfection, cells were selected using 1.7μg/ml puromycin for 48 hrs. Surviving cells were then isolated in 96-well plates by serial dilution and grown for 10 days in Serum+LIF medium. Single colonies were then trypsinized and plated into 48-well plates. Subsequently, gDNA extraction was performed in each clone and the genetic modifications were identified by PCR (methods 3.5.8.1; Table 3.12). Candidates clones with the desired deletions were then confirmed by Sanger sequencing.

Table 3.4.4. Primers used for genetic screenings.

Primer name	Sequence
	PE Sox1 CGI deletion
P1	TAATAACTGCCCTCCCCTCCCT
P2	ACAGGCCTGTTTCCTTCCTCTC
P3	CTCGAGGAAAACTGGCCGGCTG
P4	ACCTGCGGCTGGATTAATCGCT

P5	ACCAGGTTGTCCAGTGCACCCT
P6	TGGCGTTACTCGTGCTGTGCTG
PE Sox1 modules in Gata6-TAD	
P1	TGCTCAAGGGGGATTTGCAGA
P2	ATCCTGTGGGGCATGACTGT
P3	ACCGGGGTTAGCTGTTTGTT
P4	GAGCCCCTTATCCCGGTGAC
P5	CTTCGCTACCCTCCTTCTCC
P6	ACGAGTAACGCCAGCCTAGA
PE Sox1 modules in FoxA2-TAD	
P1	TGGGCTTGGGATATGCTATGCA
P2	TTCAAGACACCACTGAAGGGGG
P3	ACCGGGGTTAGCTGTTTGTT
P4	GAGCCCCTTATCCCGGTGAC
P5	CTTCGCTACCCTCCTTCTCC
P6	ATCCGGCTCCGTAGGCTGCC
PE Wnt8b modules in Gata6-TAD	
P1	TGCTCAAGGGGGATTTGCAGA
P2	ATCCTGTGGGGCATGACTGT
P3	TTTCTCCTCCAGGAATGTGG
P4	GAAGCATCCAATCCAGACCCACAATA
P5	CCTACCTCCCGGAACTCTG
P6	CTGGGTCGCTGGGAGCGCCGC
aCGI+TFBS insert in Gata6-TAD	
P1	TGCTCAAGGGGGATTTGCAGA
P2	ATCCTGTGGGGCATGACTGT
P3	ACCGGGGTTAGCTGTTTGTT
P4	GCTAGGCTGGACGTTCGGAGGT
aCGI insert in Gata6-TAD	

P1	TGCTCAAGGGGGATTTGCAGA
P2	ATCCTGTGGGGCATGACTGT
P3	AACAGGTAGGCCTAGGTGGG
P4	GCTAGGCTGGACGTTCGGAGGT
PE Sox1 modules in Gria1-TAD	
P1	TCTGCATTCCTAAGTGGGTGCT
P2	CGTGTGTATGGGGTAGGTGTGT
P3	ACCGGGGTTAGCTGTTTGTT
P4	GAGCCCCTTATCCCGGTGAC
P5	CTTCGCTACCCTCCTTCTCC
P6	ACGAGTAACGCCAGCCTAGA
pCGI in Gria1 promoter	
P1	TCTGCATTCCTAAGTGGGTGCT
P2	ACCGGGGTTAGCTGTTTGTT
P3	TGAGTGAATCCCAGGAGCCATG
P4	CTGACTCCTGCCTGGCCAGC
P5	AAGATGCTTTTGGGAATGCGGG
P6	TCAGTCAAGGCCATCCACTGTC
PE Sox1 modules in Sox7/Rp1l1-TAD	
P1	GCCTCTGGCAGTGTCATCGGTG
P2	ACCCACAGCCAGCCACTTACCT
P3	ACCGGGGTTAGCTGTTTGTT
P4	GAGCCCCTTATCCCGGTGAC
P6	ACGAGTAACGCCAGCCTAGA
PE Sox1 modules in Gria1 promoter	
P1	TGAGTGAATCCCAGGAGCCATG
P2	AAGATGCTTTTGGGAATGCGGG
P3	ACCGGGGTTAGCTGTTTGTT
P4	GAGCCCCTTATCCCGGTGAC

P6	ACGAGTAACGCCAGCCTAGA
	del36
P1	ACACCATTAGCACCCACACACT
P2	CTTGCCCCATCAGGAGATTGGA
P3	CTGTCACCTCTCTCTGCGATGT
P4	GATGCCCGTCCATTTTCTGCAA
P5	GATGGTGGGACTGGAAATGGGA
P6	CCTGTCAATCTGTCTGGGCTGA
	inv110
P1	GCCTCACTTTGAACTTCTGGCC
P2	AGCCCCCAATGTCCTTTAGCTT
P3	ACCCAGCCAAGAACTCACATCA
P4	ATTGGAGAACAGGTGGGGACAG
	inv260
P1	TGTGTCAGGGGAAAACAACA
P2	CCAACCAACTCACAGTGGAA
P3	CTGGGGCTCCCTTACTAACC
P4	TTTGCATTTGCTTGTGGACT

3.5 Molecular biology methods

3.5.1 gDNA isolation

gDNA was extracted by resuspending cells in a solution of 10 mM Tris (pH 8.0), 10 mM EDTA, 0.5% SDS and 200 µg/ml of Proteinase K. Then, cells were incubated at 55°C for 2hrs and 1 hr at 37°C with 25 µg/ml of RNAse.

3.5.2 RNA isolation, cDNA synthesis and RT-qPCR

Total RNA was isolated using Innuprep RNA kit (Table 3.4) following manufacturer´s protocol. Before cDNA synthesis, 5 µg of RNA were DNAse treated with Turbo DNAse. Then RNA was transcribed into cDNA using ProtoScript II First Strand cDNA Synthesis Kit and random primers (Table 3.5.2).

Table 3.5.2. Primers used for RT-qPCR analysis

Name	Fwd	Rev
Eef1a	TAGACGAGGCAATGTTGCTG	AGCGTAGCCAGCACTGATTT
Six3	CCTCACCCCCACACAAGTAG	CTGATGCTGGAGCCTGTTCT
Wnt8b	CAGCCATGAATCTGCACAAC	GCCGCATGATACTTCTCCTT
Sox1	TCGAGCCCTTCTCACTTGTT	CAACCCAAAAGAGCGGTAAC
Foxa2	AGCACCATTACGCCTTCAAC	CCTTGAGGTCCATTTTGTGG
Gata6	CTACACAAGCGACCACCTCA	TGTAGAGGCCGTCTTGACCT
Hptr	CAAGGGCATATCCAACAACA	GCCCCAAAATGGTTAAGGTT
Six2	GCAAGTCAGCAACTGGTTCA	CTGATGCTGGAGCCTGTTCT
Gria1	CCCGGAACTACAGGAAGCTCTC	GCGGCTGTATCCAAGACTCTCT
Sox7	CAGAACCCGGACCTGCACAACG	AGCCGCTCTGCCTCATCCACAT
Rp1l1	GGGGAGGTCAGCATCTGCCTCA	TCGTGTCTCCTGGGGTGCTGTT
Lmx1a	GACACCACCTGCTTCTACCG	ACGCATGACAAACTCATTGG
Mgst3	TAGCAAGCGGTATCGAGGAG	CGTAGCCTAAGCCTGGTCTG
Aldh9a1	TTAAGCCTTCTCCCTTCACG	TTCCGGTGAAGGAGATTTTG
Lrrc52	GGATTCCTCAGTGACCTCGT	CCTCAAAGACGTGGTGTTGG

3.5.3 ChIP-qPCR

10 million cells for histone ChIPs or 50 million cells for the rest of the ChIPs were crosslinked with 1% of formaldehyde for 10 minutes at RT and then quenched with 0.125M glycine for 10 minutes. Crosslinked cells were washed two times with PBS, resuspended in 5 ml of Lysis Buffer 1 (Table 3.2.2) and incubated 10 minutes with agitation at 4°C. Next, cells were resuspended in 5 ml of Lysis Buffer 2 (Table 3.2.2), incubated 10 minutes with agitation at RT and then resuspended in 500 µl of Lysis Buffer 3 (Table 3.2.2). The isolated nuclei were sonicated for 20 cycles for histone ChIPs or 7 cycles the other ChIPs using EpiShear sonicator at 25% amplitude. After sonication, the 10% of the chromatin solution was saved as input control and the rest of the chromatin was incubated overnight at 4°C with 1% Triton and 3 µg of antibody for histones or 10 µg of antibody for the rest of the ChIPs (Table 3.5.3.1). The next day, the chromatin was incubated for 4 hrs with 50 µl of Protein-G magnetic beads, previously blocked with 0,5% BSA. Subsequently, beads were washed 5 times with RIPA buffer (Table 3.2.2), the chromatin was eluted at 65°C with Elution buffer (Table 3.2.2) and decrosslinked, together with the input samples, overnight at 65°C. DNA was purified using ChIP DNA Clean & Concentrator (Table 3.4). ChIP samples were analyzed by qPCR using primers shown in Table 3.5.3.2.

Table 3.5.3.1 Antibodies used for ChIP-qPCR

Antibody	Company	Reference
H3K27me3	Active Motif	39155
H3K27ac	Active Motif	39133
H3K4me3	Active Motif	39159
H3K4me1	Active Motif	39297
p300	Active Motif	61401
RNAP2	Active Motif	39097
MED1	abcam	A300-793A
PHC1	Cell signalling Technologies	13768S
RING1B	Active Motif	39663
SUZ12	abcam	ab12073
CBX7	abcam	ab21873

Table 3.5.3.2. Primers used for ChIP-qPCR analysis

Name	Fwd	Rev
Chr6_neg	CTGGACTGAGGACCTTCTGC	AGGAAGGCAGATGAGGGATT
Chr2_neg	CCTGAGGCTGGAAGTTTCTG	CTCCTGGGATTAAAGGCACA
Gata6_promoter	CCGCACCCAGCAGCTTGTAGAG	TCTGCCGGAAAACTGCAGCCTG
Gria1_promoter	GCCCCCTTTGGAGAAAGCACCC	AACCCATCACCCCTGCCTCTCC
Gata6_TAD P1	CCTAATTCAGTTAGGTGACTGC	CCTTGAGCATAGTCAAGCTTTC
Gata6_TAD P2	TCAAGGGGAAGACAGAATTTGA	TATCTTCTTGGAGGCAGTCACA
Gria1_TAD P1	TAGAATGGGATGGGGGTTGAGC	ACAAACACCCCAATGCCAACAG
Gria1_TAD P2	TCCCAAGGCAAATGATTGGTCA	TGGAGCTGTTTTGTCCTCTAAGCT
PE_Sox1	ACCTGCGGCTGGATTAATCGCT	GAAGCTTCCGTGTTCCCACCCG

3.5.4 Immunofluorescence

mESC and AntNPC were grown in microscopic slides and fixed for 10 minutes with 3,7% paraformaldehyde at RT. Fixed cells were permeabilized with 0.1% Triton X-100 for 15 minutes and blocked in PBS with 5% BSA for 1 hour at RT. Next, cells were incubated with primary antibodies (Table 3.16) in 5% BSA overnight at 4°C. The next day, cells were rinsed with PBS and incubated with secondary antibodies (Table 3.16) in 5% BSA for 30 minutes at RT. Nuclei were stained with DAPI for 10 minutes at RT and cells were mounted with Mounting medium.

Table 3.5.4. Antibodies used for Immunofluorescence

Antibody	Company	Reference
GATA6	R&D systems	AF1700
SOX1	R&D systems	AF3369
anti-Goat Alexa Fluor 488	Invitrogen	A32814
anti-Goat Alexa Fluor 594	Invitrogen	A32758

3.5.5 Bisulfite sequencing

400 ng of gDNA was bisulfite converted using the EZ DNA Methlation Kit (Table 3.4) following manufacturer instructions. The sequence of interest were then amplified by PCR using EpiTaq polymerase (xxx) using primers described in Table XXX. Finally, PCR products ere clones into pGEM-T vector (Table xxx) and sequenced with the M13 reverse primer.

Table 3.5.5. Primers used for bisulfite-converted sequences.

Name	Fwd	Rev
Gata6 TAD_BS	TTTGATTATGTTTAAGGAGTAGGTTG	TCAACTTTATAAAACCATTAAATACAC
Gria1 TAD_BS	TTTTTTTGTTTGTTAGTATTTAAAAAG	TCAAAAATAATACATAATTTCAAAAATAAT

3.5.6 Formaldehyde-assisted isolation of regulatory elements (FAIRE)

10 M cells were crosslinked with 1% of formaldehyde for 10 minutes at RT and lysated with subsquent washes with Lysis Buffer 1, 2 and 3 (Table xxx). Then, chromatin was sonicated for 4 cycles at 25% amplitude using EpiShear Sonicator. 10% of the total chromatin fragments were saved as input control and the rest was enriched for regulatory elements by three rounds of phenol/chloroform purification followed by ethanol precipitation (Giresi et al., 2007). The FAIRE and input DNAs were analysed by qPCR using two negative controls and primers P1 and P2 from Gata6-TAD and Gria1-TAD (Table X).

3.5.7 Circular chromatin conformation capture (4C-seq)

4C-seq was performed following a previous described protocol (Stadhouders et al., 2013) with minor modifications. Briefly, 10 M cells were crosslinked with 1% of formaldehyde for 20 minutes at RT and quenched with 0.125M glycine for 10 minutes. Then, cells were washed with PBS and incubated in 4C Lysis Buffer (Table X) for 10 minutes on ice. Nuclei were then resuspended in 500 µl of 1.2X Cutsmart restriction buffer (NEB) with 0.3% SDS and incubated at 37°C for 1 hour. Triston x-100 was added to a concentration of 2% and incuated for 1 hour at 37°C. Chromatin was then digested overnight at 37°C with 400 U of NlaIII restriction enzyme. The next day, NlaII was inactivated with SDS to a concentration of 1.6% and incubated for 20 minutes at 65°C. The chromatin fragments were then resuspended in 6 ml of 1X 4C ligation buffer plus 1% Triton x-100, incubated for 1 hour at 37°C while shaking at 300 rpm and then ligated with 100 U of T4 ligase for 4 hours at 16°C. Ligated chromatin was treated with RNAse A for 45 minutes at 37°C and then de-crosslinked with 300 mg of Proteinase K at 75°C overnight. DNA was then purified by phenol/chloroform extraction and ethanol purification, resuspended in 100 µl of water and digested overnight with 50 U of DpnII at 37°C. The next day, DNA was purified by phenol/chloroform extraction and ethanol purification, resuspended in 15 ml of water and ligated overnight with 200 U of T4 ligase at 16°C. DNA samples were then

subjected once more to phenol/chloroform extraction and ethanol purification, resuspended in 100 ul of water and purified by column.

4C DNA was amplified by 4C inverse PCR (xxx) using primers located within the selected viewpoints (Table X) and sent for Illumina sequencing to the Erasmus Medical Center in Rotterdam. Resulting reads were mapped and analyzed with r3C-seq to identify contacts between the selected viewpoints and the rest of the genome.

3.5.8 PCRs

3.5.8.1 PCR for genotyping mESC clones

In order to determine the genotype of CRISPR-engineered mESC clones, gDNA was extracted with QuickExtract DNA Extraction Solution (Table xxx) and subjected to a PCR with GoTaq polymerase using the following conditions:

Component	Volume
5X Green buffer	2.5 µl
dNTPs 10mM each	0.25 µl
MgCl2 25mM	2 µl
gDNA	2 µl
Primer Fwd 100 µM	0.125
Primer Rev 100 µM	0.125
GoTaq	0.07
H2O	10 µl

95°C – 2minutes
95°C – 30 seconds
59°C – 30 seconds
72°C – 1 minute/kb
} 30 cycles
72°C – 5 minutes

3.5.8.2 PCR for amplification of bisulfite-converted DNA

PCR amplification of bisulfite-treated template DNA was performed using EpiTaq polymerase under the following conditions:

Component	Volume
EpiTaq	12.5 µl
10X Buffer	0.75 µl
MgCl2	0.75 µl
dNTPs	1 µl
gDNA (50 ng/µl)	1 µl
Primer Fwd (100 µM)	0.25 µl
Primer Rev (100 µM)	0.25 µl
H2O	to 25 µl

95°C – 10 seconds
55°C – 30 seconds
72°C – 30 seconds
} 35 cycles

3.5.8.3 PCR for amplification of homology arms and PE modules

In order to amplify the homology arms used for CRISPR insertions, KAPA PCR was performed in WT (E14) mESC gDNA using the following conditions:

Component	Volume
2X KAPA Master Mix	12.5 µl
100µM primer fwd	0.75 µl
100µM primer rev	0.75 µl
gDNA 100 ng/µl	1 µl
H2O	to 25 µl

95°C – 3 minutes
98°C – 20 seconds
68°C – 15 seconds
72°C – 15 seconds
} 35 cycles
72°C – 1 minute

3.5.8.4 PCR for amplification of donor template

Donor template was amplified from the cassette-vector (xxx) using the KAPA HiFI PCR system under the following conditions:

Component	Volume
2X KAPA Master Mix	12.5 µl
100µM primer fwd	0.75 µl
100µM primer rev	0.75 µl
vector 1 ng/µl	1 µl
H2O	to 25 µl

95°C – 3 minutes
98°C – 20 seconds
70°C – 15 seconds
72°C – 30 seconds/kb
} 35 cycles
72°C – 1 minute

3.5.8.5 4C PCR

4C DNA was amplified by inverse PCR using primers located in the viewpoint loci as previously described (Stadhouders et al., 2013). Briefly, for each vewipoint a primer p5 was designed in close proximity to the NlaIII restriction site and a primer p7 was designed close to the DpnII restriction site. Illumina adaptors were added in the 5` end of both primers together with a specific barcode sequence for each sample. The PCR was performed with the Expand Long Template PCR system (Table xxx) using the following conditions:

Component	Volume
10X PCR buffer	5 µl
dNTPs 10mM each	1 µl
35 pmol/µl p5 primer	1 µl
35 pmol/µl p5 primer	1 µl
Expand Long polymerase	0.7
4C DNA 50ng/µl	1 µl
GoTaq	0.07
H2O	To 50 µl

94°C – 2minutes
94°C – 10 seconds
60°C – 1 minute
68°C – 3 minutes
} 30 cycles
68°C – 5 minutes

Table 3.5.8.5. Primers used for 4C DNA amplification

Name	Viewpoint	Sequence (adaptor+barcode+primer)
p5 WT	*Gata6* promoter	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGCCAATGGACTTACTAAAAGGGTAGACTGTC
p5 TFBS	*Gata6* promoter	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGGGACTTACTAAAAGGGTAGACTGTC
p5 CGI	*Gata6* promoter	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGTGAAAGGACTTACTAAAAGGGTAGACTGTC
p5 TFBS+CGI	*Gata6* promoter	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTCTTGTAGGACTTACTAAAAGGGTAGACTGTC
p7	*Gata6* promoter	CAAGCAGAAGACGGCATACGAGGACTTACTAAAAGGGTAGACTGTC
p5 WT	*Gata6*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGCCAATCTGTTAGCTCTCTCTACTGGCA
p5 TFBS	*Gata6*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGCTGTTAGCTCTCTCTACTGGCA
p5 CGI	*Gata6*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGTGAAACTGTTAGCTCTCTCTACTGGCA
p5 TFBS+CGI	*Gata6*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTCTTGTACTGTTAGCTCTCTCTACTGGCA
p7	*Gata6*-TAD insertion site	CAAGCAGAAGACGGCATACGATCGATGGCACTGTTATTCTGAC
p5 WT	*Gria1*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGCCAATCAGACTACATCATAACAAGGAGCA
p5 TFBS	*Gria1*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGCAGACTACATCATAACAAGGAGCA
p5 CGI	*Gria1*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGTGAAACAGACTACATCATAACAAGGAGCA
p5 TFBS+CGI	*Gria1*-TAD insertion site	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTCTTGTACAGACTACATCATAACAAGGAGCA
p7	*Gria1*-TAD insertion site	CAAGCAGAAGACGGCATACGATCGAGTATGACCAATCATTTGCC
p5 WT	*PE Six3*	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGATTAACACCCAAACGCCTGC
p5 del36	*PE Six3*	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTGTGAAAATTAACACCCAAACGCCTGC
p5 inv	*PE Six3*	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTCTTGTAATTAACACCCAAACGCCTGC
p7	*PE Six3*	CAAGCAGAAGACGGCATACGAAATGTAAGTGGAGCTCGTACTG
p5 WT	*PE Sox1*	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGCTCTCCAGCAGCGAAAGCAAA
p5 PE Sox1-/-	*PE Sox1*	AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTC TTCCGATCTGTGAAACTCTCCAGCAGCGAAAGCAAA
p7	*PE Sox1*	CAAGCAGAAGACGGCATACGAAATGTAAGTGGAGCTCGTACTG

3.6 Statistical analysis

3.6.1 RT-qPCR analysis

Gene expression levels were calculated using the 2ΔCt method and normalize to the average signal of two housekeeping genes (*Eef1a* and *Hptr* from Table 3.5.2). Standard deviations were calculated from technical triplicates reactions.

3.6.2 ChIP-qPCR analysis

ChIP levels were calculated as percentage of input and normalized to the average signal of two negative regions (Chr2 and Chr6 from Table 3.5.3.2). Standard deviations were calculated from technical triplicates reactions and represented as error bars.

3.6.3 Hi-C analysis

The Hi-C analysis was done in collaboration with Victor Sanchez-Gaya. He did the computational analysis and I contributed to the design and interpretation of the data.

For the analysis, we used the .hic files for two Hi-C replicates (GSM3752487, GSM3752488) generated in ESCs, and the .cool format files from GSE98671 (GSM2644945, GSM2644946) at a 20-kb matrix resolution. The .hic files were converted to .cool format using a 5-kb matrix resolution with the hic2cool software (https://github.com/4dn-dcic/hic2cool). For both datasets the corresponding replicates in .cool format were merged and normalized (Abdennur & Mirny, 2020).

When a PE-distal was found in a TAD with a devTSS that contain a CAP-CGI within <3 kb, both coordinates were selected to define a PE–gene pair of the group “Developmental genes with CGI+ promoters”. On the other hand, when a PE-distal was found in a TAD with TSS that do not

contain a CAP-CGI within <3 kb, both coordinates were selected to define a PE–gene pair of the group "Genes with CGI- promoters). In addition, we applied the following filters: (i) PE–gene pairs were balanced to compare groups of PE–gene pairs without significant differences in their linear genomic sizes. PE–gene pairs with similar lengths were selected by applying the nearest neighbor matching method (without replacement and ratio=1) using MatchIt (https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf) and considering Group "Developmental genes with CGI+ promoters" as the treatment condition; (ii) only TSSs of genes with expression <1 FPKM were considered. We used TADs defined in mESC_Dixon2012-raw_TADs.txt (Dixon et al., 2012).

The pile-up plots for the GSE130723 and GSE98671 Hi-C datasets were generated with coolpup.pyc106 using a padding of ±50 kb or ±100 kb, respectively.

3.6.4 TF motif analysis

The TF motif analysis was done in collaboration with Victor Sanchez-Gaya. He did the computational analysis and I contributed to the design and interpretation of the data.

The genomic coordinates of PEs and AEs were defined by p300 peaks identified in ESCs (Cruz-Molina et al., 2017) and located >2.5 kb away from any RefSeq TSS. Among the PEs, we considered only those with a CAP-CGI within <3 kb and that did not overlap with the p300 peaks defining the PEs. Motif analyses were performed separately for the CAP-CGIs and the p300 peaks associated with the selected PEs using Homer (Heinz et al., 2010) (using as parameters: -size given -mset vertebrates) and SeqPos (Liu et al., 2011) (using a mouse and human curated motif database and *de novo* motif searches)

3.6.5 Whole-genome bisulfite sequencing (WGBS) analysis

The WGBS analysis was done in collaboration with Tore Bleckwehl. He did the computational analysis and I contributed to the design and interpretation of the data.

For WGBS analyses we used public data from ESCs cultured with MEK and GSK3 inhibitors (2i) (Habibi et al., 2013), epiblast-like cells (EpiLCs) (Zylicz et al., 2015), epiblast stem cells (EpiSCs) (Zylicz et al., 2015), serum + LIF ESCs (GSE82125) and neural progenitor cells (GSE82125). The adapters were trimmed with Trim Galore and mapped to the mm10 reference genome using Bismark-v.0.16.1 (Krueger & Andrews, 2011) and bowtie2 v.2.2.9 (Langmead & Salzberg, 2012). For each cell type, the CpG methylation levels were estimated with the Bismark methylation extractor, considering only CpGs with a coverage of 3–100 reads. For visualization of CpG methylation levels around pCGIs and oCGIs, the average CpG methylation signal was visualized with deeptools v.3.3.1 (Ramírez et al., 2016).

3.6.6 ChIP-seq profile plots

The ChIP-seq profile plots were done in collaboration with Victor Sanchez-Gaya. He did the computational analysis and I contributed to the design and interpretation of the data.

PE-distal were separated into four groups: (i) PEs with overlapping TFBS/p300 and CAP-CGIs; (ii) PEs with TFBS/p300 separated by 1 bp to 1 kb from a CAP-CGI; (iii) PEs with TFBS/p300 separated by 1–3 kb from a CAP-CGI; and (iv) PEs without CAP-CGIs within 3 kb. The coordinates of AEs without CAP-CGI in <3 kb were also considered. mESC datasets from GSE157748 and GSE89209 (H3K27me3), GSE104067 (TET1) and GSE126862 (KDM2B) were used. Profiles plots were generated using computeMatrix and plotProfile from deepTools (Ramírez et al., 2016).

3.6.7 Genetic analysis of oCGIs

The genetic analysis of oCGIs was done in collaboration with Victor Sanchez-Gaya. He did the computational analysis and I contributed to the design and interpretation of the data.

For the analysis, we used PE coordinates from (Cruz-Molina et al., 2017) and filtered the list for PEs that were more than 2.5kb away from any TSS (PE-all). We sub-filtered the list for PEs that were at least 10kb away from any TSS (PE-distal). CAP-CGI coordinates were obtained from (R. S. Illingworth et al., 2010). NMI coordinates were obtained from (Long, Sims, et al., 2013).

In mESCs, developmental genes commonly contain broad H3K27me3 domains. Therefore, to

A relevant feature for many important cell-identity genes is that they are frequently embedded within broad domains of H3K27me3 when they are not expressed (Rehimi et al., 2016). Therefore, we processed H3K27me3 ChIP-Seq and its corresponding input data generated in mESC (GSE89209; H3K27me3 ID: SRR4453259, Input ID: SRR4453262). Next, H3K27me3 peaks with respect to the Input were called with MACS2102 using the broad peak calling mode. Only those peaks with a fold-enrichment > 3 and q value < 0.1 were maintained. Subsequently, peaks closer than 1kb were merged using bedtools, and associated with a protein coding gene if they overlapped a TSS. Lastly, the size distribution of the H3K27me3 peaks associated with genes was studied and developmental genes were defined as those genes with the largest peak sizes. To do so, the knee of the size distribution was determined with findiplist() (inflection R package; [https://cran.r-project.org/web/packages/inflection/vignettes/inflection.html]). Upon curvature analysis, a threshold of 6 Kb was defined and, thus, all genes with a H3K27me3 peak length larger than 6kb were considered as developmental genes (devTSS).

CAP-CGIs were associated with PE-distal or devTSS if located less than 3kb away from them. In addition, to get a representation of the bulk genome composition, a third group of random regions was created. To create this random group, each of the regions associated with a PE-distal

was randomly relocated along the genome 1000 times, maintaining its size. Finally the set of all randomized regions constituted the random group.

To retrieve the DNA sequences of the studied regions, the BSgenome package was used (Pagès, 2020), using as reference the unmasked mouse mm9 genome. For each region: its length, the percentage in G+C, the percentage in CpG and the CpG observed/expected ratio was calculated. The %CpG was calculated as the ratio of CpG dinucleotide counts with respect to half the total region length. The ratio of observed to expected CpG is calculated according to the formula presented in (Gardiner-Garden & Frommer, 1987).

4- RESULTS

4.1 Genes with CGI-rich promoters respond to PEs

To investigate the rules dictating the compatibility between enhancers and genes, we implemented a genetic engineering approach consisting on insertions of a PE into different TADs (Figure 4.1.1). By selecting TADs whose genes are inactive and devoid of active enhancers both in mESC and AntNPC, we reasoned that any transcriptional effect in these genes should only be attributed to the PE insertion. To implement this approach, we used a CRISPR-Cas9 knock-in system (Yao et al., 2018) to insert the *PE Sox1(+35)* at approximately 100kb from *Gata6*-, *FoxA2*-, *Gria1*-TSS (Figure 4.1.2). The inserted *PE Sox1(+35)* sequence contains a highly conserved TFBS and a nearby oCGI. The three loci were selected because they contain genes with different promoter types according to a previous classification (Lenhard et al., 2012). As PE-target genes, *Gata6* and *FoxA2* contain type III promoters, which correspond to developmental genes with large CGI extending to the body of the gene (Figure 4.1.2). In contrast, *Gria1* contain a type I promoter, which is a promoter without CGIs (Figure 4.1.2). Using this strategy, we obtained two homozygous clones for each insertion site (Figure 4.1.3). To study the transcriptional effect of the inserted PE, we did RT-qPCR in mESC and AntNPC engineered cell lines and measured the expression of *Gata6*, *FoxA2* or *Gria1*. In addition, *Sox1* expression was measured as a positive control for proper differentiation. As expected, the insertion of the *PE Sox1(+35)* did not affected *Gata6, FoxA2* or *Gria1* expression in mESC (Figure 4.1.4A-D). However, in AntNPC, *Gata6* or *FoxA2* expression was induced by ~50- or ~13-fold with respect to WT cells, respective (Figure 4.1.4A-D). In contrast, the *PE Sox1(+35)* insertion did not had any transcriptional effect on *Gria1*-TAD engineered AntNPC. This result suggest that some genes are responsive to PEs, albeit they are normally expressed in a different

cellular states than the one in which the PE is active. Interestingly, not all the genes respond to PEs, which also suggest that there might be genetic rules regulating PE responsiveness.

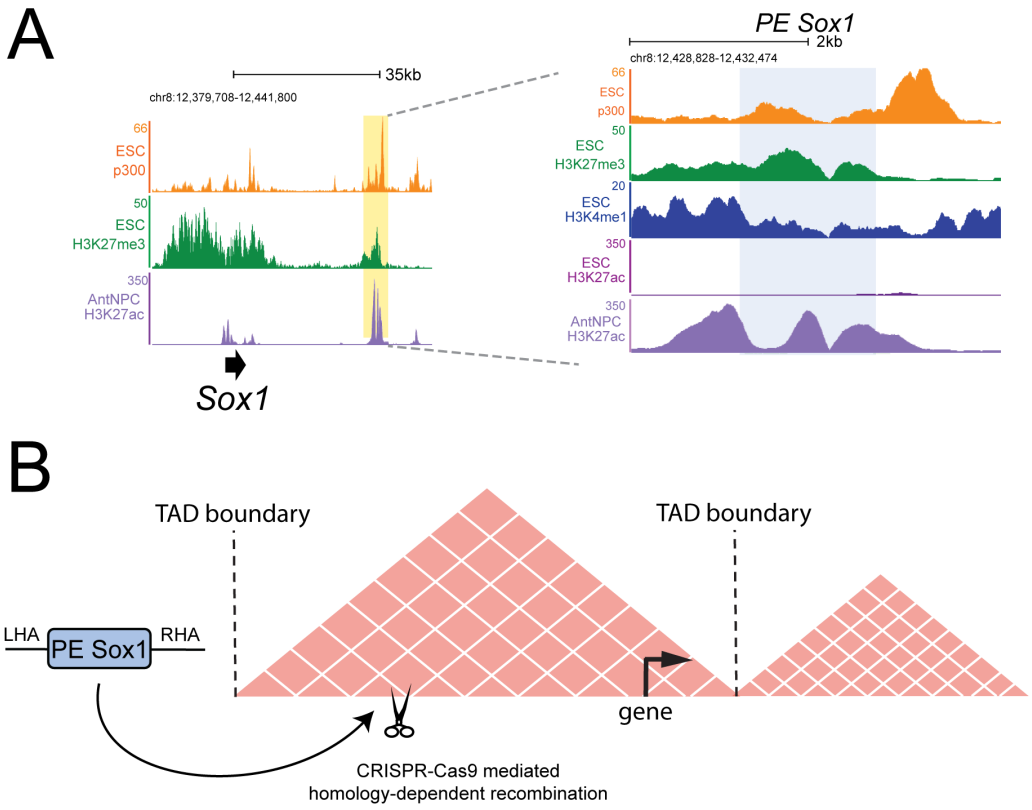


Figure 4.1.1. Knock-in of PEs in mESC. **(A)** The left panel shows ChIP-seq data (Cruz-Molina et al., 2017) from mESC (P300, H3K27me3) and AntNPC (H3K27ac) at the *Sox1* locus. *PE Sox1(+35)* is highlighted in yellow. The right left panel shows, highlighted in blue, the *PE Sox1(+35)* sequence used for knock-in insertions and its associated ChIP-seq (Cruz-Molina et al., 2017) data from mESC (p300, H3K27me3, H3K4me1, H3K27ac) and AntNPC (H3K27ac). **(B)** Strategy used to insert *PE Sox1(+35)* at different TADs. TAD boundaries are represented as dotted lines. A PCR product containing the *PE Sox1(+35)* sequence and 300 bp homology arms (LHA and RHA) specific for each targetted locus was used for as template for CRISPR-Cas9 mediated homology-dependent recombination.

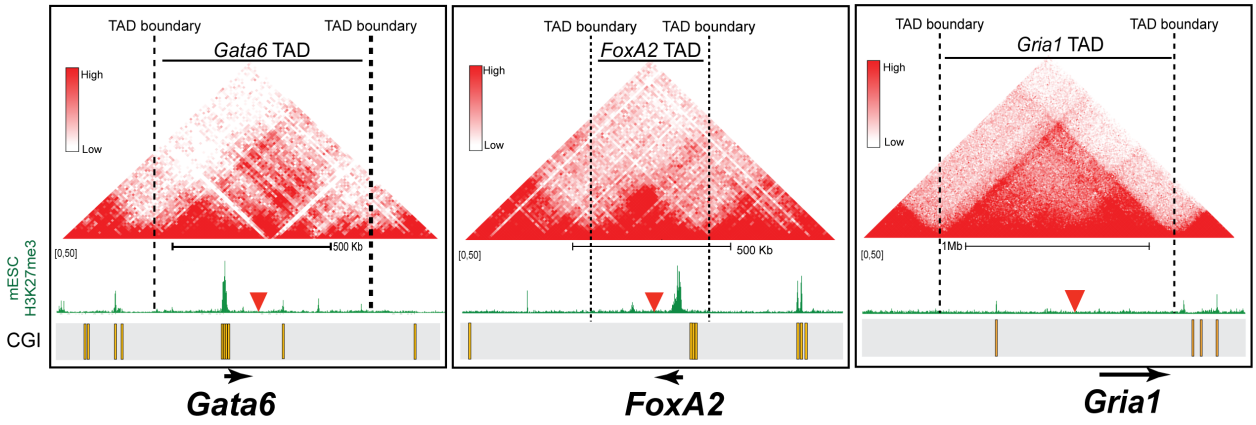


Figure 4.1.2. Selected TADs for Knock-in of PEs in mESC. Hi-C profiles (Bonev et al., 2017) of the *Gata6*-, *FoxA2*- and *Gria1*-TAD. TAD boundaries are denoted with dotted lines. H3K27me3 ChIP-seq signals in mESC are shown in green (Cruz-Molina et al., 2017) and CGIs are indicated as yellow rectangles. The red triangle indicates the integration site of the inserts.

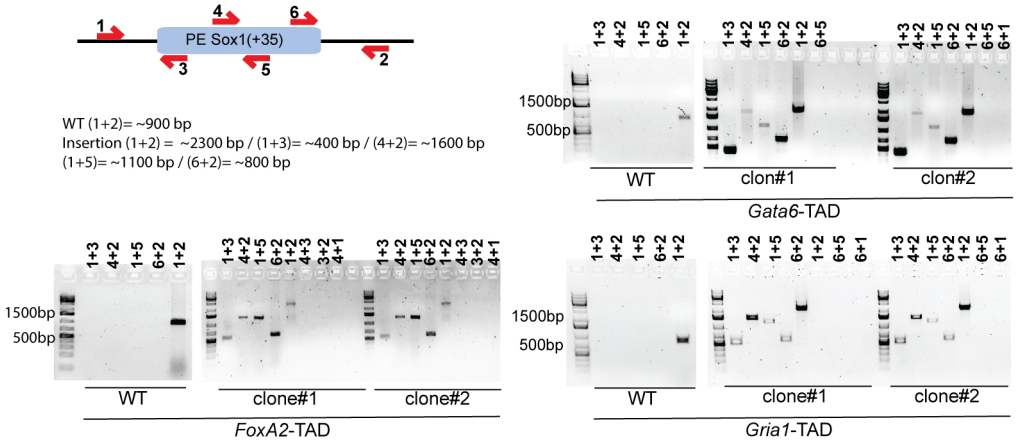


Figure 4.1.3. Genotyping of inserted PEs in mESC. For the identification of the *PE Sox1(+35)* insertions, primer pairs flanking the insertion borders (1+3, 4+2, 1+5 and 6+2), amplifying potential duplications (3+2 and 6+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used. The PCR results obtained for WT ESCs and for two ESC clonal lines with homozygous insertions of the *PE Sox1(+35)* in the different TADs (i.e. *Gata6*-, *FoxA2*- and *Gria1*-TAD) are shown.

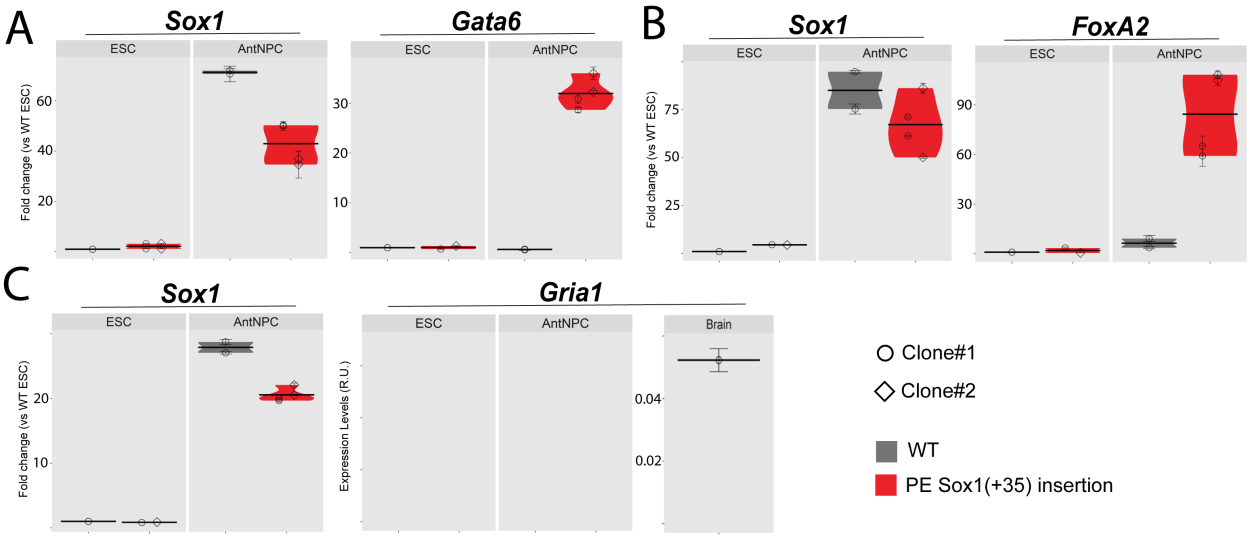


Figure 4.1.4. Not all genes are responsive to PE insertions. RT-qPCR experiments showing the expression levels of *Sox1* and *Gata6* (A), *FoxA2* (B) or *Gria1* (C) in mESC and AntNPC cell lines that were either WT (grey) or homozygous for the insertion of *PE Sox1(+35)* in the *Gata6*- (A), *FoxA2*-TAD (B) or *Gria1*-TAD (C) (red). For the cells with the PE insertions, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*). Gria1 expression values are presented as arbitrary units (R.U.) since it was not detectable (ND) except in the mouse adult brain.

The above result suggest that genes with CGI-rich promoters (i.e. Gata6 and FoxA2) are able to respond to PEs (i.e. *Gata6* and *FoxA2*), in contrast to genes with CGI-poor promoters (Lenhard et al., 2012). To experimentally assess whether developmental genes with large CGI clusters in their promoters are particularly responsive to PEs, we inserted the *PE Sox1(+35)* into the *Sox7/Rp1l1*-TAD, between *Sox7* and *Rp1l1* (24 kb from Sox7 and Rp1l1 TSSs) (Figure 4.1.5A and B). *Sox7* and *Rp1l1* are both inactive in ESCs and AntNPCs, but differ in their type of promoter: *Sox7* contain a type III promoter (i.e. a CGI cluster that extends from its promoter region to part of its 5′ coding region), while *Rp1l1* contain a type I promoter (i.e. a CGI-poor

promoter). Differentiation of this cell line into AntNPC and subsequent RT-qPCR assay showed that *PE Sox1(+35)* is not able to induce Rp1l1 (Figure 4.1.5C). In contrast, *Sox7* was strongly induced by the PE insertion. Altogether, these results suggest that genes containing CGI-rich promoters might be particularly responsive to PEs.

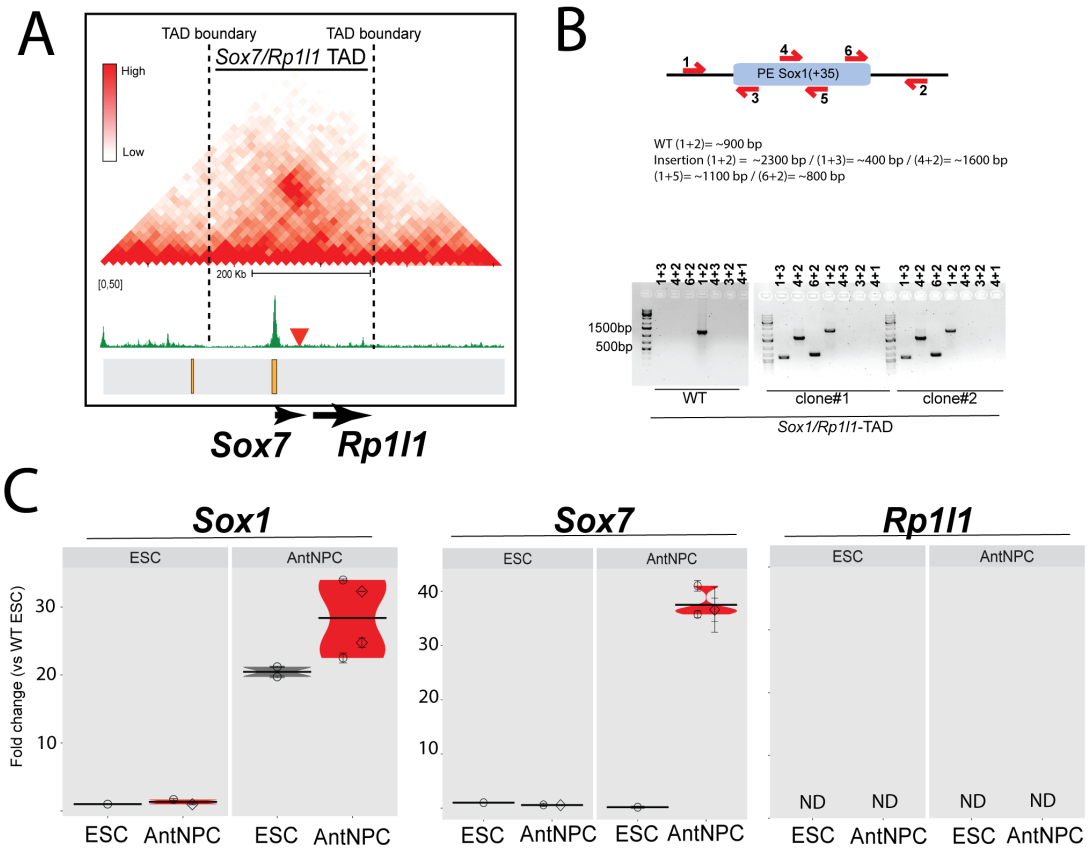


Figure 4.1.5. CGI-rich gene, but not CGI-poor, respond to PE induction. (A) Hi-C profiles (Bonev et al., 2017) of the *Sox7/Rp1l1*-TAD. TAD boundaries are denoted with dotted lines. H3K27me3 ChIP-seq signals (Cruz-Molina et al., 2017) in mESC are shown in green and CGIs are indicated as yellow rectangles. The red triangle indicates the integration site of the insert. **(B)** For the identification of the *PE*

Sox1(+35) insertions, primer pairs flanking the insertion borders (1+3, 4+2, 1+5 and 6+2), amplifying potential duplications (3+2 and 6+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used. The PCR results obtained for WT ESCs and for two ESC clonal lines with homozygous insertion of the *PE Sox1(+35)* in the *Sox7/Rp1l1*-TAD are shown. **(C)** RT-qPCR experiments showing the expression levels of *Sox1, Sox7* and *Rp1l1* in mESC and AntNPC cell lines that were either WT (grey) or homozygous for the insertion of *PE Sox1(+35)* in the *Sox7/Rp1l1*-TAD (red). For the cells with the PE insertions, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*).

4.2 Role of CGIs in PE-target genes pre-looping

Before becoming active, PEs are already bookmarked in ESCs with preformed contacts with their target genes (Cruz-Molina et al., 2017). To further investigate these preformed topologies between PE and genes containing CGIs, we analysed Hi-C public data from mESC to assess the frequency of contacts between PEs and genes that are CGI-rich or CGI-poor. Interestingly, PEs commonly establish physical contacts with CGI-rich promoters from genes located their same TAD, in contrast to genes containing CGI-poor promoters (Figure 4.2.1A). To experimentally test this correlation, we did 4C-seq experiments in *Gata6*- and *Gria1*-TAD engineered mESC using as a viewpoint the insertion site or the target gene promoter (i.e. *Gata6* or *Gria1* promoter). We observed a strong interaction between the *PE Sox1(+35)* inserted at the *Gata6*-TAD and the *Gata6* promoter, but no interaction between the *PE Sox1(+35)* inserted at the *Gria1*-TAD and the *Gria1* promoter (Figure 4.2.1B).

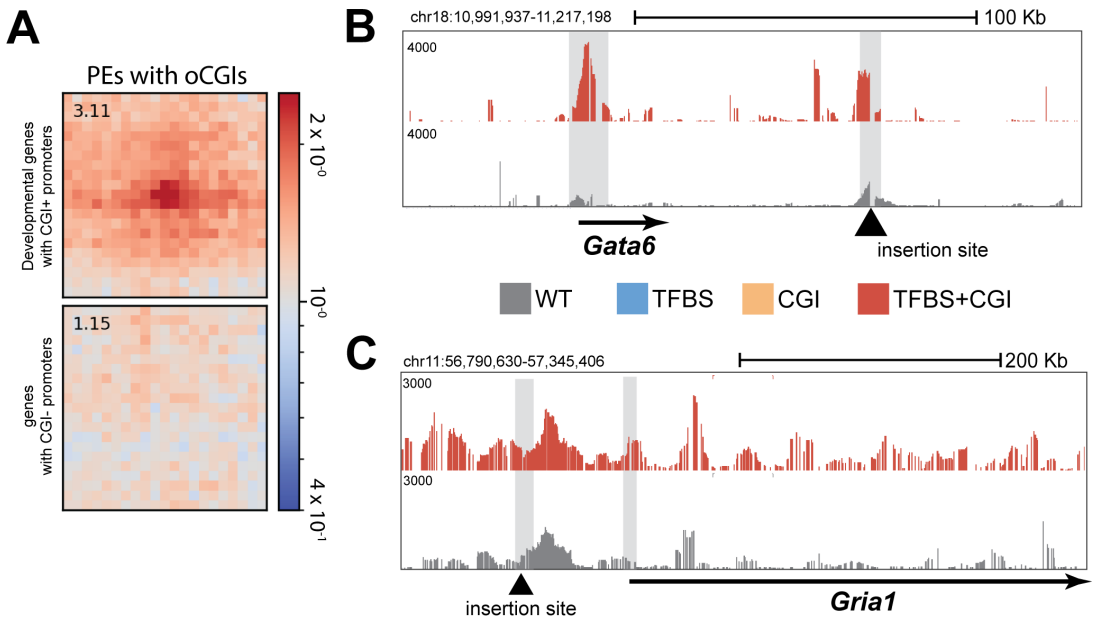


Figure 4.2.1. CGI-rich genes physically interact with PEs located at their same TAD. (A) Hi-C data previously generated in ESCs was used to generate pile-up plots showing the average interaction between PEs and either developmental genes with CGI-rich promoters (upper panel) or genes with CGI-poor promoters (lower panel). For each of the considered PE-gene pairs, both the PE and the gene were located within the same TAD. **(B and C)** 4C-seq experiments were performed using the *Gata*6- **(B)** or the *Gria1*-TAD insertion site **(C)** as viewpoints in ESCs that were either WT (grey) or homozygous for the insertions of the *PE Sox1(+35)* (red) within *Gata*6- **(B)** or the *Gria1*-TAD insertion site **(C)**.

Our lab has previously suggested that the pre-looping between PEs and their target genes is dependent on PcG complexes located at both sites (Crispatzu et al., 2021; Cruz-Molina et al., 2017). Therefore, we reasoned that impairing PcG binding at a PE should abolish its interaction with the target gene. Since CGI act as PcG responsive elements in vertebrate genomes, we tested this hypothesis by using CRISPR-Cas9 to delete the endogenous CGI of *PE Sox1(+35)* (*PE Sox1(+35)CGI*$^{-/-}$; Figure 4.2.2A). ChIP-qPCR for the H3K27me3 mark in WT mESC and in two *PE Sox1(+35)CGI* $^{-/-}$ clones showed that the deletion of the CGI severely reduce H3K27me3

levels at the *PE Sox1(+35)* (Figure 4.2.2B). To study the effect of the CGI deletion in the looping between *PE Sox1(+35)* and *Sox1*, we did 4C-seq experiments in WT mESC and *PE Sox1(+35)CGI*$^{-/-}$. In agreement with our hypothesis, the deletion of *PE Sox1(+35)CGI* resulted in reduced physical contacts with the *Sox1* promoter (Figure 4.2.2C).

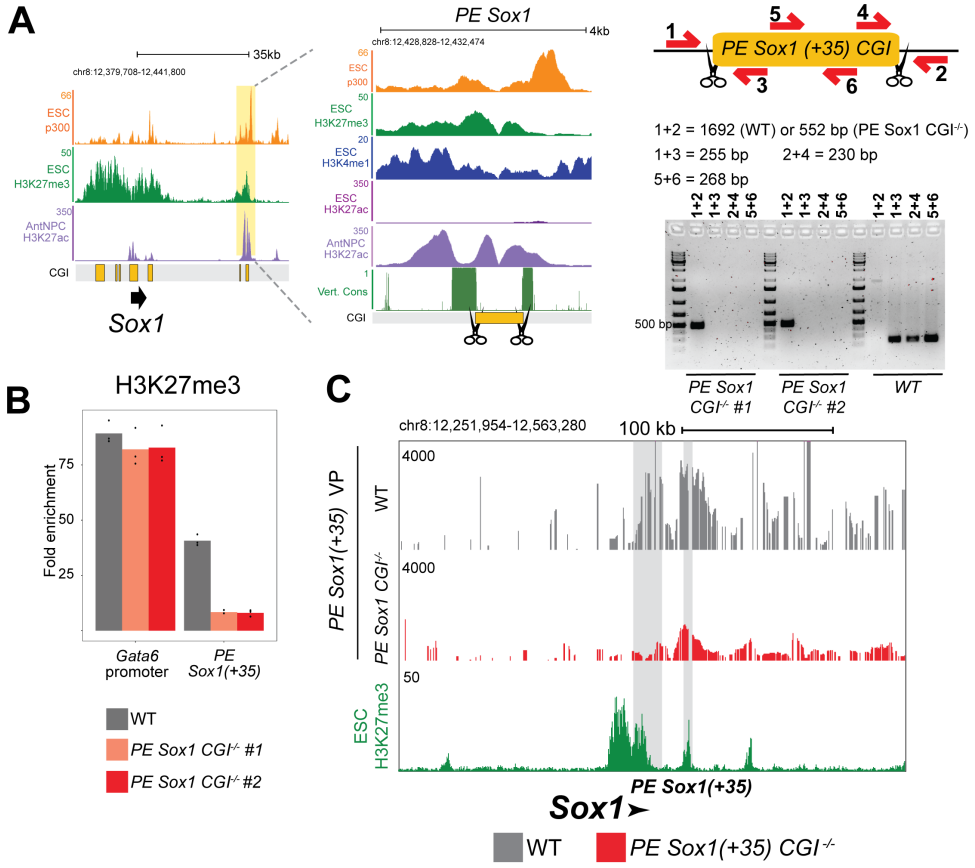


Figure 4.2.2. CGIs recruit PcG at PEs and regulate their physical interactions with target genes. (A) The left panel shows ChIP-seq data from ESCs (p300 and H3K27me3) and AntNPCs (H3K27ac) at the *Sox1* locus (Cruz-Molina et al., 2017). The *PE Sox1(+35)* is highlighted in yellow. CGIs are represented as yellow rectangles. The right panel shows a close-up view of the *PE Sox1(+35)* with additional epigenomic and genomic data. The represented CGIs correspond to those computationally defined in the UCSC browser according to the following criteria: GC content > 50%; Length > 200 bp; CpG Observed to expected ratio > 0.6. Vert. Cons.= vertebrate PhastCons. The right panel shows the PCR screening for detecting the *PE Sox1(+35)CGI* deletion. Primer pairs flanking each of the deletion breakpoints (1+3 and 4+2), located within the deleted region (5+6) or amplifying a large or small fragment depending on the

absence or presence of the deletion (1+2) were used. The PCR results obtained for WT ESCs and for two ESC clonal lines with homozygous deletions of the PE *Sox1(+35)CGI* (*PE Sox1(+35)CGI*$^{-/-}$) are shown. **(B)** H3K27me3 levels at *PE Sox1(+35)* were measured by ChIP-qPCR in WT ESCs (gray), and in two *PE Sox1(+35)CGI*$^{-/-}$ ESCs clones using primers adjacent to the deleted region. ChIP-qPCR signals were normalized against two negative regions. Error bars correspond to standard deviations from three technical replicates. **(C)** 4C-seq experiments were performed using the *PE Sox1(+35)* as a viewpoint in ESCs that were either WT (gray) or homozygous for the deletion of *PE Sox1(+35) CGI* (*PE Sox1 CGI*$^{-/-}$ *(*red*))*. H3K27me3 ChIP-seq signals in ESCs are shown in green. The genomic location of *PE Sox1(+35)* and *Sox1* are highlighted in gray.

In order to further study the contribution of CGIs to the topological function of PEs, we now dissected PE *Sox1(+35)* into two elements (i.e. the TFBS and the CGI) and inserted them alone or together into the same TADs studied above (i.e. *Gata6*-, *FoxA2*-, *Gria1*- and *Sox7/Rp1l1*-TAD) (Figure 4.4.1A). We assume that this approach will be more adequate to study the effect of CGI in individual PEs than loss-of-function approaches (i.e. deletion of CGIs), because of two main reasons: (i) CGI are close to and commonly overlapping TFBS, which complicates the deletion of specific PE components; (ii) other regulatory elements (i.e. redundant enhancers) present in the target gene landscape could mask the regulatory function of CGIs. Using this strategy, we obtained two homozygous clones for each insertion (i.e *PE Sox1(+35)TFBS, Sox1(+35)CGI* or *PE Sox1(+35)TFBS+CGI)* in each locus (i.e. *Gata6*-, *FoxA2*-, *Gria1*- and *Sox7/Rp1l1*-TAD) (Figure 4.4.1B).

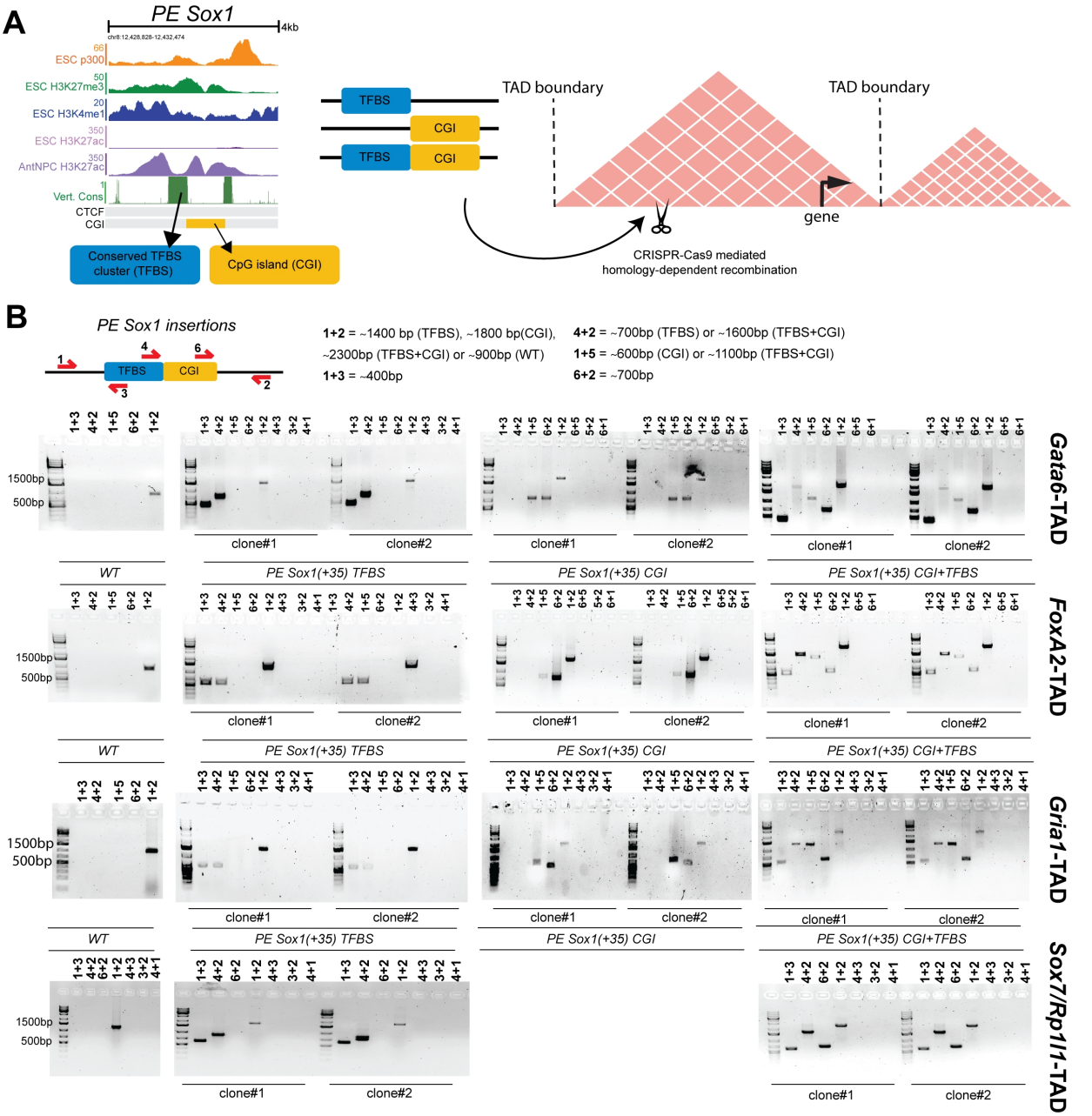


Figure 4.4.1. Generation of cell lines with knock-in of different PEs modules within different TADs. **(A)** The left panel shows ChIP-seq data from ESCs (p300 and H3K27me3) and AntNPCs (H3K27ac) at the *Sox1* locus (Cruz-Molina et al., 2017). The *PE Sox1(+35)* is highlighted in yellow. CGIs are represented as yellow rectangles. The right panel shows a close-up view of the *PE Sox1(+35)* with additional epigenomic and genomic data. The represented CGIs correspond to those computationally

defined in the UCSC browser according to the following criteria: GC content > 50%; Length > 200 bp; CpG Observed to expected ratio > 0.6. Vert. Cons.= vertebrate PhastCons. **(B)** For the identification of the *PE Sox1(+35)* insertions, primer pairs flanking the insertion borders (1+3, 4+2, 1+5 and 6+2), amplifying potential duplications (3+2 and 6+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used. The PCR results obtained for WT ESCs and for two ESC clonal lines with homozygous insertion of each of the *PE Sox1(+35)* insertions (TFBS, CGI and TFBS+CGI) in the *Gata6*-, *FoxA2*-, *Gria1*- or *Sox7/Rp1l1*-TAD are shown.

Next, to evaluate whether the inserted CGI was able to recruit PcG, we performed ChIP assays for PcG associated histone marks (i.e. H3K27me3 and H2AK199ub) and PcG subunits (i.e. SUZ12 and RING1B) in ESC lines containing the different *PE Sox1(+35)* components within the *Gata6*- or *Gria1*-TAD (Figure 4.4.2). PRC2 associated mark H3K27me3 and its subunit SUZ12 were enriched around all the inserts containing the CGI (*PE Sox1(+35)CGI* and *PE Sox1(+35)TFBS+CGI* in *Gata6*- and *Gria1*-TAD) (Figure 4.4.2). Intriguingly, PRC1 associated mark H2K119ub and its subunit RING1B were more around the *Sox1(+35)TFBS+CGI* insert than in the *PE Sox1(+35)CGI* in both the *Gata6*- and *Gria1*-TAD (Figure 4.4.2). This suggest that the binding of PRC1 is increased by the combination of TFBS and oCGIs. Since previous research suggest that PRC1, but not PRC2, is involved in PcG-mediated long-range contacts (Isono et al., 2013), we hypothesized that only the combination of TFBS+CGI and a CGI-rich promoter would generate a loop. To test this, we did 4C-seq experiments in *Gata6*-and *Gria1*-TAD engineered mESC using as a viewpoint the insertion site or the target gene promoter (i.e. *Gata6* or *Gria1* promoter). Indeed, we observed a strong interaction between the *PE Sox1(+35)TFBS+CGI* and the *Gata6* promoter, while the *PE Sox1(+35)TFBS* or the *PE Sox1(+35)CGI* did not show any specific interaction with this gene (Figure 4.4.3A). Contrary, neither the *Sox1(+35)TFBS+CGI* nor the other components inserted in the *Gria1*-TAD showed interaction with *Gria1* (Figure 4.4.3B). Together, these results suggest that CGI at PEs are necessary for PcG loading and pre-looping between PEs and their target genes.

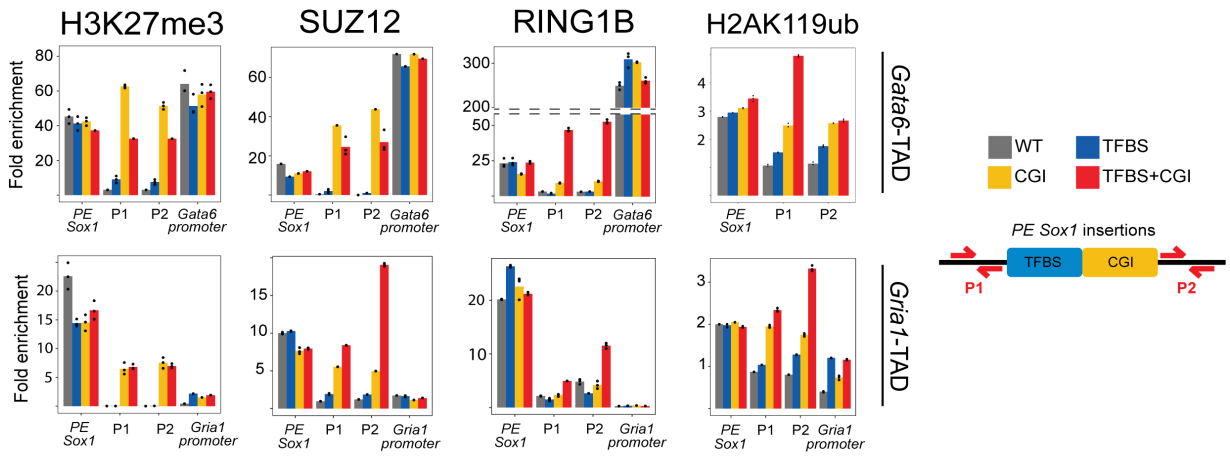


Figure 4.4.2. oCGIs are necessary for the recruitment of PcG to PEs. PRC1 (H3K27me3 and SUZ12) and PRC2 (RING1B and H2AK119ub) levels at *PE Sox1(+35)*, the insertion site (P1 and P2) or the target gene promoter (*Gata6* or *Gria1* promoter) were measured by ChIP-qPCR in WT ESCs (gray), and cell lines with the different *PE Sox1(+35)* insertions at the Gata6- (upper panel) or Gria1-TAD (lower panel) (TFBS (blue), CGI (yellow) and TFBS+CGI (red)). ChIP-qPCR signals were normalized against two negative regions.

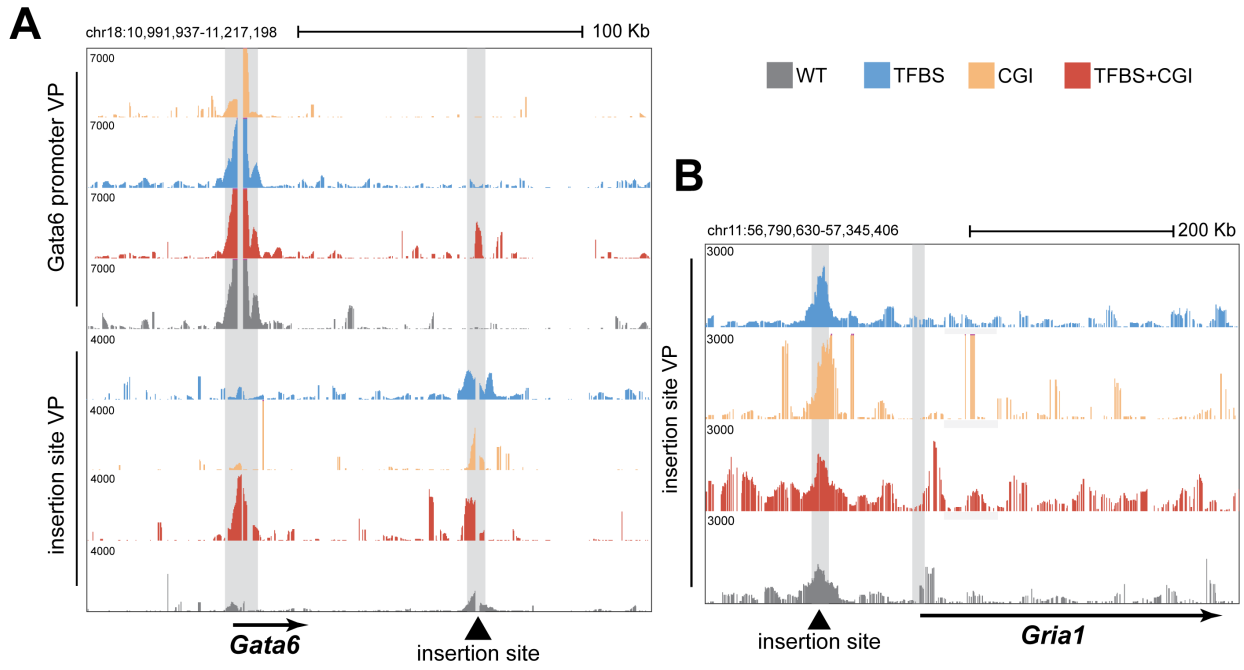


Figure 4.4.3. oCGIs regulate the looping between PEs and their target genes in ESC. (A) 4C-seq experiments were performed using the *Gata6* promoter (upper panel) or the insertion site as viewpoints in ESCs that were either WT (grey) or homozygous for different insertions at the *Gata6*-TAD (TFBS (blue), CGI (yellow) and TFBS+CGI (red)). **(B)** 4C-seq experiments were performed using the *Gria1* promoter as viewpoint in ESCs that were either WT (grey) or homozygous for different insertions at the *Gria1*-TAD (TFBS (blue), CGI (yellow) and TFBS+CGI (red))

4.3 Role of CGIs in PE-target gene activation

Our lab previously suggested that pre-looping between PEs and their target genes might be important for the cis-activation capacity of the enhancer (Crispatzu et al., 2021; Cruz-Molina et al., 2017). Therefore, we reason that PE-associated CGIs can influence the induction of target genes, by at least regulating their topological configuration. Therefore, to investigate the regulatory role of CGIs in PEs, we differentiated WT mESC, two *PE Sox1(+35)CGI* $^{-/-}$ clones and one cell line where the complete *Sox1(+35)* was deleted (*PE Sox1(+35)*$^{-/-}$) (Cruz-Molina et al., 2017) into AntNPC and measured *Sox1* expression by RT-qPCR. In mESC, the deletion of the CGI did not affected *Sox1* expression (Figure 4.3.1A). In AntNPC, *PE Sox1(+35)CGI* $^{-/-}$ and

PE Sox1(+35)$^{-/-}$ clones showed a reduced *Sox1* induction by >2- and >4-fold compared to WT, respectively (4.3.1A). This result suggest indeed a positive role of PE-associated CGI to the cis-activation capacity of the enhancer.

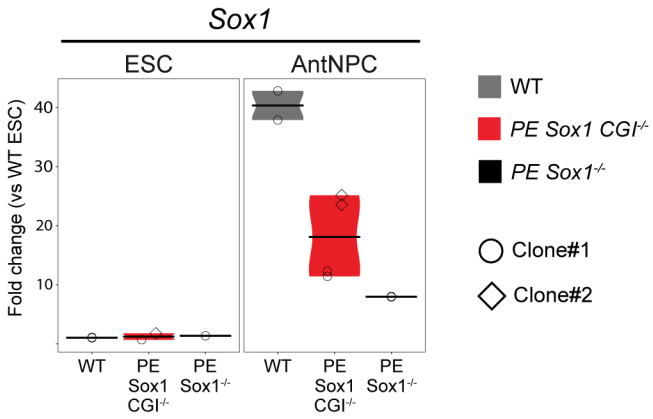


Figure 4.3.1. oCGIs have regulatory functions in PEs. The expression of *Sox1* (was measured by RT-qPCR in ESCs (left panels) and AntNPCs (right panels) that were either WT (grey), homozygous for the deletion of the *PE Sox1(+35)CGI (PE Sox1(+35)CGI*$^{-/-}$*)* or homozygous for the deletion of the *PE Sox1(+35) (PE Sox1(+35)*$^{-/-}$*)*. For *PE Sox1(+35)CGI*$^{-/-}$, two different clonal cell lines (circles and diamonds) were studied. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average and standard deviation (error bars) from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*) and are presented as fold-changes with respect to WT ESCs.

To validate these results, we studied the transcriptional effect of CGIs in the engineered PE sequences by doing RT-qPCR in mESC and AntNPC with the different PE modules insertions (i.e. *PE Sox1(+35)TFBS*, *PE Sox1(+35)CGI* and *PE Sox1(+35)TFBS+CGI)* in *Gata6-, FoxA2-, Gria1-* or *Sox7/Rp1l1-TAD* and measure the expression of the respective genes. In addition, *Sox1* expression was measured as a control. As expected, none of the *PE Sox1(+35)* insertions induced gene expression in mESC (Figure 4.3.2A-D). However, in AntNPC, the *Sox1(+35)TFBS+CGI* insertion strongly induced the expression of CGI-rich genes (i.e. *Gata6, FoxA2* and *Sox7*) with respect to WT cells, but did not affected the transcription of CGI-poor genes (i.e. *Gria1* and

Rp1l1) (4.3.2A-D). In contrast, the *Sox1(+35)TFBS* insertion only led to a weak induction of *Gata6* or *Sox7* (4.3.2A-D). Lastly, the *PE Sox1(+35)CGI* had no effect in any of the AntNPC engineered cell lines (4.3.2A-D).

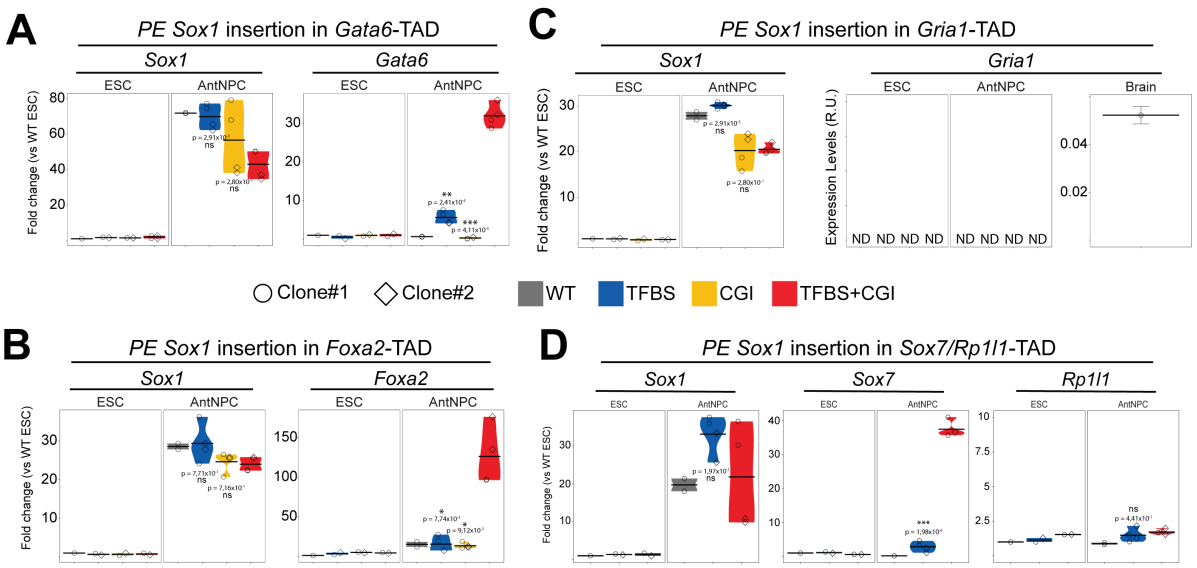


Figure 4.3.2. oCGIs boost the cis-activation capacity of PEs. (A-D) The expression of *Gata6* **(A)**, *Foxa2* **(B)**, *Sox1* **(A-D)**, *Gria1* **(C)**, Sox7 **(D)** *and Rp1l1* **(D)** was measured by RT-qPCR in ESCs (left panels) and AntNPCs (right panels) that were either WT (grey) or homozygous for the insertions of the different *PE Sox1(+35)* modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)) in the *Gata6*- **(A)**, *Foxa2*- **(B)**, *Gria1*- **(C)** or *Sox7/Rp1l1*-TAD **(D)**. For the cells with the PE module insertions, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average and standard deviation (error bars) from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*) and are presented as fold-changes with respect to WT ESCs. The expression differences between AntNPCs with the TFBS+CGI module and either WT AntNPCs or AntNPCs with the different PE modules were calculated using non-paired t-tests (***: fold-change>2 & p<0.0001; ns: not significant; fold-change<2 or p>0.01).

Finally, to asses if the previous results are consistent with other PEs, we generated new cell lines containing modules from the *PE Wnt8b(+21)* (i.e. *PE Wnt8b(+21)TFBS*, *PE Wnt8b(+21)CGI* and *PE Wnt8b(+21)TFBS+CGI)* in the *Gata6*-TAD (Figure 4.3.2A). Like *Sox1*, *Wnt8b* is associated to a PE (i.e. *PE Wnt8b(+21)),* which controls *Wnt8b* expression in AntNPC (Cruz-Molina et al., 2017). RT-qPCR experiments in mESC and AntNPC of these engineered cell lines showed again that the combination of TFBS and CGI (i.e. *Wnt8b(+21)TFBS+CGI)* strongly induce *Gata6* in AntNPC, while the TFBS and CGI modules do not affect *Wnt8b* expression Figure 4.3.2B).

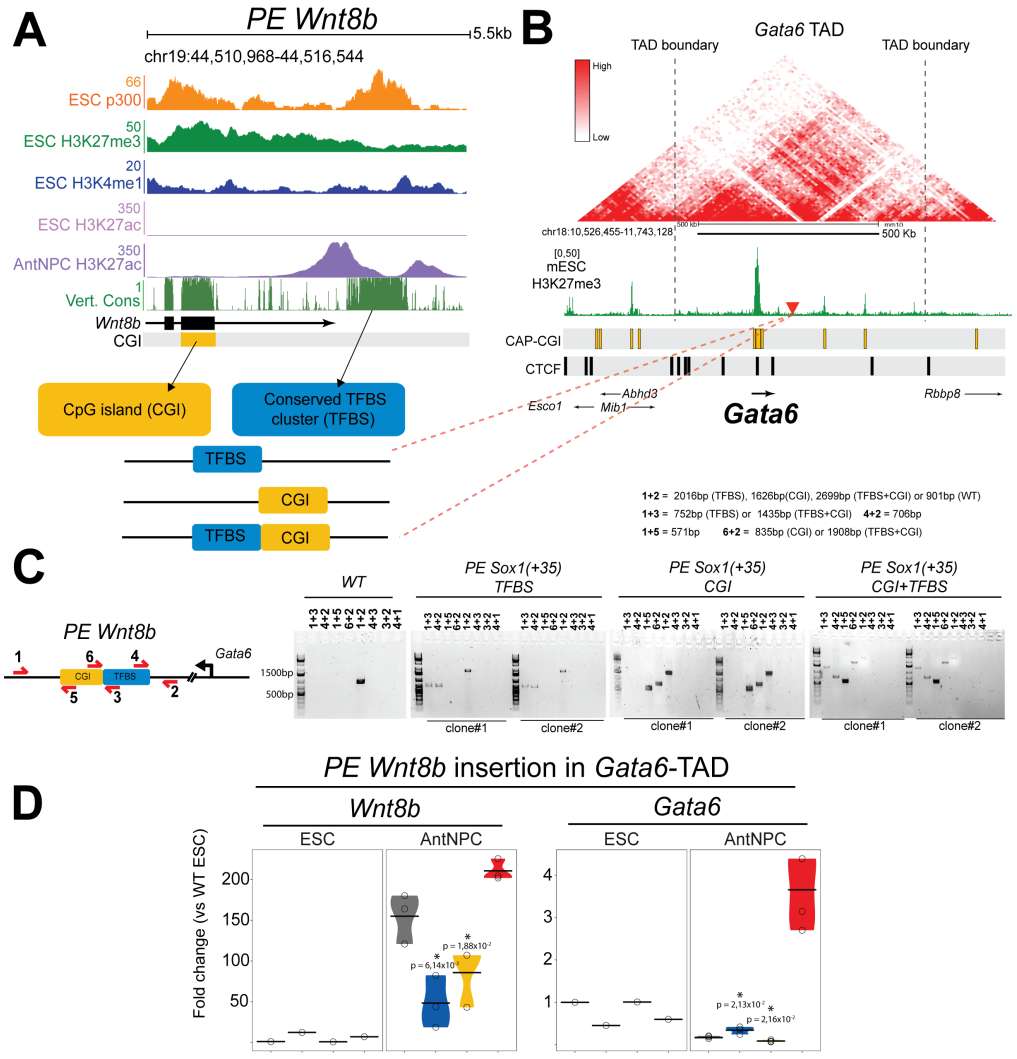


Figure 4.3.2. Modular knock-in of *PE Wnt8b(+21)* within the *Gata6-TAD*. ChIP-seq data (Cruz-Molina et al., 2017) from mESC (P300, H3K27me3, H3K4me, H3K27ac) and AntNPC (H3K27ac) at the *PE Wnt8b(+21)* locus. The represented CGIs correspond to those computationally defined in the UCSC browser according to the following criteria: GC content > 50%; Length > 200 bp; CpG Observed to expected ratio > 0.6. Vert. Cons.= vertebrate PhastCons. The lower left panel shows the combinations of *PE Wnt8b(+21)* elements inserted into the *Gata6*-TAD (i.e. (i) *PE Wnt8b(+21)TFBS;* (ii) *PE Wnt8b(+21)CGI;* (iii) *PE Wnt8b(+21)TFBS+CGI*). **(B)** Hi-C profiles of the *Gata6*-TAD (Bonev et al., 2017). TAD boundaries are denoted with dotted lines. H3K27me3 ChIP-seq signals in mESC are shown in green (Cruz-Molina et al., 2017) and CTCF binding sites in ESC are shown as black rectangles. CGIs

are indicated as yellow rectangles. The red triangle indicates the integration site of the *PE Wnt8b(+21)* modules. **(C)** *PE Wnt8b(+21)* modules insertions at the *Gata6*-TAD were evaluated by PCR using primers flanking the insertion borders (1+3 and 4+2; 1+5 and 6+2; or 1+3 and 6+2), amplifying potential duplications (4+3, 3+2 and 4+1; or 6+5, 5+2 and 6+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used. **(D)** RT-qPCR experiments showing the expression levels of *Wnt8b* and *Gata6* in mESC and AntNPC cell lines that were either WT or have the different *PE Wnt8b(+21)* insertions (i.e. TFBS, CGI, TFBS+CGI) in the *Gata6*-TAD. For the cells with the PE module insertions, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*).

4.4 Genetic and epigenetic properties of CGIs associated to PEs

Having identified a role for PE-associated CGI as topological and transcriptional regulators of PEs, we decided to study the genetic and epigenetic properties of CGIs associated to PEs. First, to identify CGI associated to PEs, we used coordinates from CAP-identified CGIs (CAP-CGIs) (R. S. Illingworth et al., 2010), computational-identified CGIs (cCGIs) (Gardiner-Garden & Frommer, 1987) or Bio-CAP-identified CGIs (non-methylated islands (NMIs)) (Long, Sims, et al., 2013) and overlapped them with previously determined PE coordinates (Cruz-Molina et al., 2017). CAP-CGIs and NMIs were included in the analysis since computational models are not a precise methods for identifying promoter-distally located CGI, also referred as orphan CGIs (oCGIs) (Long, Sims, et al., 2013). We found that over 60% of PEs are within 3 Kb of biochemically identified oCGIs (i.e. CAP-CGIS and NMIs), while this percentage is significantly lower for computational-identified CGIs (Figure 4.4.1A). As stated before, this could be explained by the fact that this algorithm underestimates the abundance of CGI, especially those that are distal to TSS (Long, Sims, et al., 2013). In addition, we found that PEs are commonly associated with only 1 CAP-CGI, while developmental genes usually contain clusters of CAP-CGIs in their promoters (Figure 4.4.1B). Next, we genetically compared CAP-CGIs associated to

PE (PE-CAP-CGIs) with those found in proximity of developmental gene TSSs (devTSS-CAP-CGIs) and random regions. In line with previous characterization of oCGIs (Bell & Vertino, 2017), PE-CAP-CGIs are shorter in length and present a lower CpG density than devTSS-CAP-CGIs (Figure 4.4.1C and D). Here, I we will use the term oCGI for cCGIs, CAP-CGIs and NMIs that are located at least 3kb from a PE.

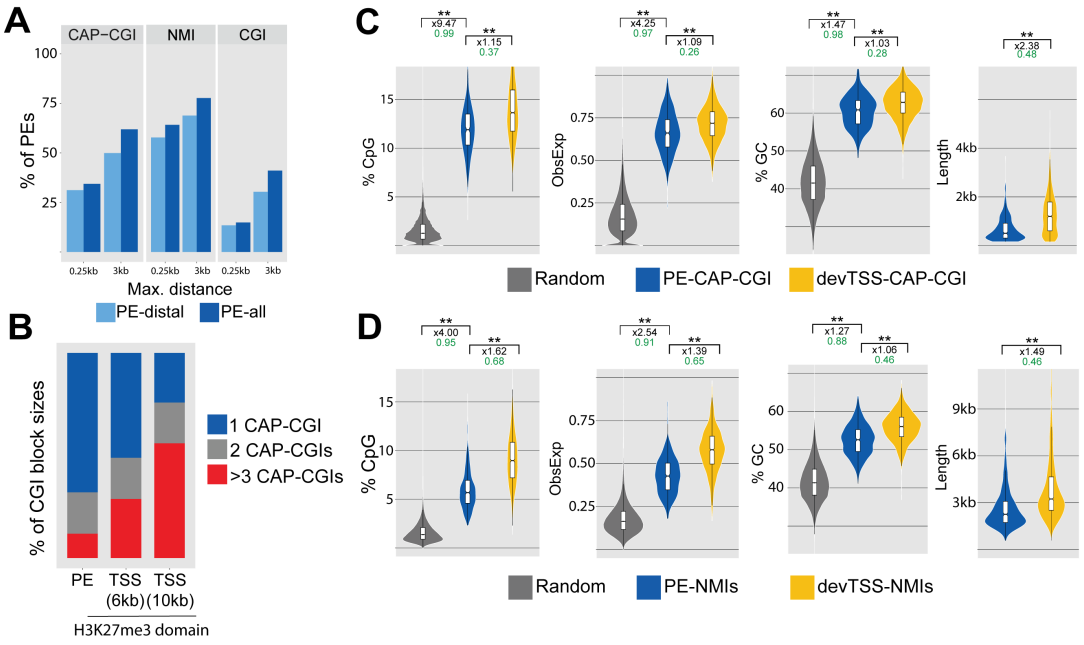


Figure 4.4.1. oCGIs associated to PEs are a genetically distinct subgroup of CGIs. (A) Percentage of PEs that are found within 0.25 or 3 Kb from: a CGI identified by CAP assay (R. S. Illingworth et al., 2010) (CAP-CGI), a CGI identified by Bio-CAP assay (Long, Sims, et al., 2013) (NMIs) or a computationally-defined CGI (Gardiner-Garden & Frommer, 1987). PE list was filtered to include only PEs that are at least 10 Kb from TSS (light blue, PE-distal). **(B)** Percentage of CAP-CGI block sizes with 1 (blue), 2 (grey) or more than 3 CAP-CGIs (red) found within 3 Kb from a PE-distal or a TSS with at least 1 CAP-CGI. TSS list was subdivided into two groups: TSS with H3K27me3 domains of more than 6 Kb or 10 Kb of length. **(C and D)** Comparison of the CpG percentage, observed/expected CpG ratio, GC percentage and sequence length between random regions and biochemically identified CGIs associated to PEs or to TSS of genes with at least 1 CGI. On top of each plot, the asterisks indicate P-values calculated using unpaired Wilcoxon tests with Bonferroni correction for multiple testing (** = p.val < $1e^{-10}$; * p.val <

0.05); the numbers in black indicate the median fold changes between the indicated groups; the green numbers indicate non-negligible Cliff Delta effect sizes. CAP-identified CGIs coordinates (-CAP-CGI) were used in **(C)** and Bio-CAP identified CGIs coordinates (NMIs) were used in **(D)**.

Our data suggest that oCGI boost the TFBS cis-activation capacity, although they are not able to induce the expression of the target genes when they are inserted alone. This suggest that the boosting capacity of oCGIs is not attributed to binding sites for TFs. We test did by doing an in silico motif analysis using Homer and SeqPos for CAP-CGIs or TFBS (defined by p300 enrichment peaks) from PEs in which both elements do not overlap (Figure 4.4.2). This analysis showed a high enrichment for pluripotency TFs motifs in TFBS but not specific enrichment for any factors in CAP-CGIs, which were mostly enriched in CpG-rich motifs instead.

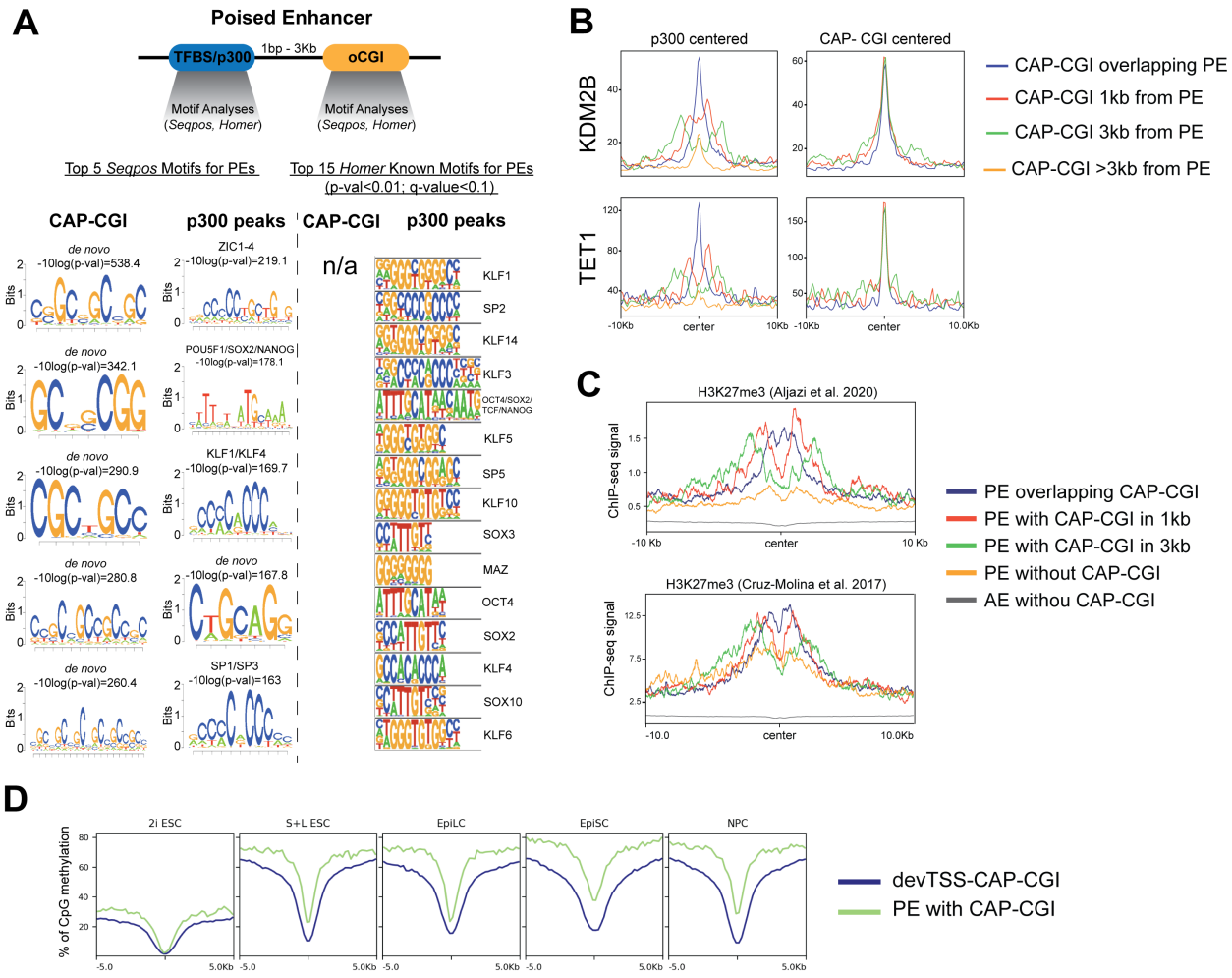


Figure 4.4.2. oCGIs associated to PEs do not contain TF cognate motifs but serve as docking site for CxxC-domian containing proteins. (A) TF motif analyses using Homer and Seqpos for PEs with a CAP-CGI located in less than 3kb and that do not overlap with the p300 peaks defining the PEs (Cruz-Molina et al., 2017). Motif analyses were performed separately for the CAP-CGIs and the p300 peaks associated with the selected PEs. (B) ChIP-seq levels for KDM2B (upper panel) and TET1 (lower panel) are shown around: (I) PEs with overlapping TFBS/p300 peaks and CAP-CGIs, (ii) PEs with TFBS/p300 peaks separated by 1bp-1kb from CAP-CGIs, (iii) PEs with TFBS/p300 peaks separated by 1-3kb from CAP-CGIs and (iv) PEs without CAP-CGIs within 3kb. ChIP-seq profile plots were generated using either the p300 peaks of the PEs (left) or the CAP-CGIs (right) associated with the PEs as midpoints. **(C)** ChIP-seq levels of H3K27me3 around: PE with p300 peaks overlapping CAP-CGI (blue); PE with p300 peaks separated by 1bp to 1 kb from CAP-CGIs (red); PE with p300 peaks separated by 1 to 3 kb (green); PE without CAP-CGI within 3kb from its p300 peak (orange); Active enhancers without CAP-CGI within 3kb from its p300 peak (violet). **(D)** Percentage of CpG methylation around CAP-CGI associated to PE

(blue) or CAP-CGI associated with developmental gene TSS (green) in different cellular states: 2i cultured mESC (2i ESC), Serum+LIF cultured mESC (S+L ESC), Epiblast like cells (EpiLC), Epiblast stem cells (EpiSC) and neural progenitor cells (NPC).

The genetic features of promoter-associated CGIs (pCGIs) allow them to be bound by proteins containing CxxC-domains (e.g. KDM2B, TET1), which also have additional domains and/or are part of complexes that can modify chromatin (e.g. PcGs). Consequently, pCGI are typically hypomethylated and enriched in PcG associated histone marks (e.g. H3K27me3). Having showed that oCGI might not contain TF motifs, we reasoned that the boosting capacity of oCGIs could be attributed to its CG-rich density. To investigate if oCGI also bind proteins containing ZF-CxxC domains, we used public ChIP-seq data sets to measure the levels of KDM2B and TET1 in subsets of PEs which contain CAP-CGIs at different distances from their central region (i.e. the highest P300 ChIP-seq peak signal): (i) PE overlapping CAP-CGIs; (ii) PEs with CAP-CGIs between 0 and 1kb; (iii) PEs with CAP-CGIs between 1 and 3kb and (iv) PEs without CAP-CGIs in their proximities. Both KDM2B and TET are highly enriched around the center of the CAP-CGI in all the subset of PEs, but their binding is reduced at the center of the PE when the CAP-CGI is more distal (Figure 4.4.2B). Interestingly, H3K27me3 ChiP-seq data shows that even those PE with CAP-CGI at 3kb are highly enriched in PcG binding (Figure 4.4.2C). In addition, analysis of publically available whole genome bisulfite data demonstrate that oCGIs share a similar hypomethylation pattern as pCGIs (Figure 4.4.2D).

Together, these results suggest that the high density of CpGs might be the required genetic information necessary for effecting their regulatory role. To investigate this, we designed an artificial CGI (aCGI) and inserted it alone or together with the *PE Sox1(+35)TFBS* at the *Gata6*-TAD (*PE Sox1(+35)TFBS+aCGI;* Figure 4.4.3A). Notably, RT-qPCR experiments in AntNPC showed that *Sox1(+35)TFBS+aCGI* cells expressed considerably higher levels of *Gata6* than *Sox1(+35)TFBS* cells (Figure 4.4.3B), whereas the insertion of the aCGI alone did not induced

Gata6 expression (Figure 4.4.3C). In conclusion, this data suggest that the CpG-richness of oCGIs might be sufficient to boost the cis-activation capacity of PEs.

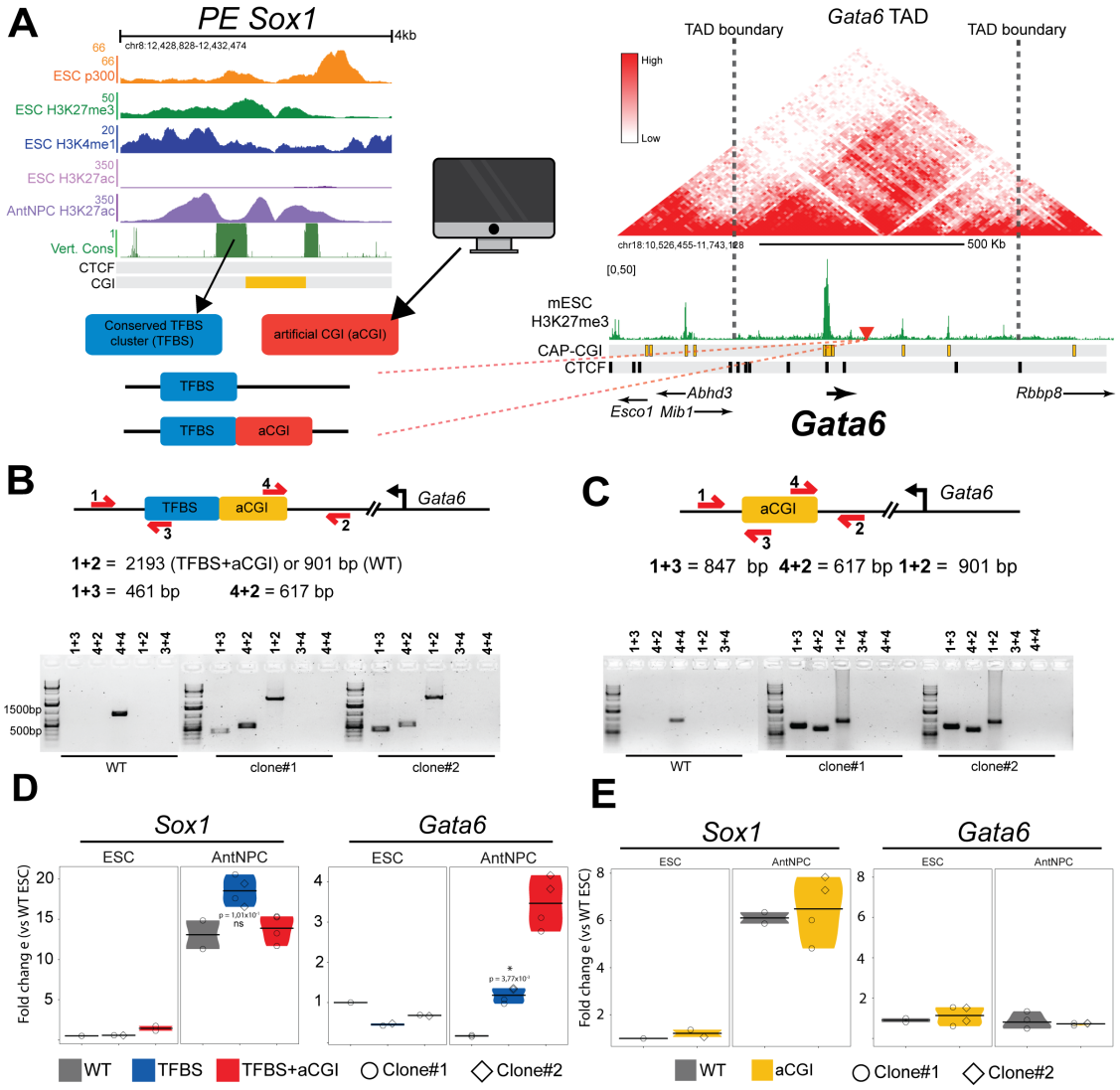


Figure 4.7. Modular knock-in of an artificial oCGI and the *PE Sox1(+35)* within the *Gata6-TAD*. The upper left panel shows ChIP-seq data (Cruz-Molina et al., 2017) from mESC (P300, H3K27me3, H3K4me, H3K27ac) and AntNPC (H3K27ac) at the *PE Sox1(+35)* locus. The represented CGIs correspond to those computationally defined in the UCSC browser according to the following criteria: GC content > 50%; Length > 200 bp; CpG Observed to expected ratio > 0.6. Vert. Cons.= vertebrate PhastCons. The lower left panel shows the elements inserted into the *Gata6*-TAD (i.e. *PE Sox1(+35)TFBS* alone or together with the artificial oCGI (*PE Sox1(+35)TFBS+aCGI*)). The right panel shows Hi-C profiles of the *Gata6*-TAD (Bonev et al., 2017). TAD boundaries are denoted with dotted lines. H3K27me3 ChIP-seq signals in mESC are shown in green (Cruz-Molina et al., 2017) and CBS in ESC[116] are shown as black rectangles. CGIs are indicated as yellow rectangles. The red triangle indicates the integration site of the inserts. **(B and C)** Insertions of the *PE Sox1(+35)TFBS* and *PE Sox1(+35)TFBS+aCGI* (B) or the artificial oCGI alone (aCGI) (C) were evaluated by PCR using primers flanking the insertion borders (1+3 and 4+2), amplifying potential duplications (4+3, 3+2 and 4+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used. **(D and E)** RT-qPCR experiments showing the expression levels of *Sox1* and *Gata6* in mESC and AntNPC cell lines that were either WT or have the following insertions: *PE Sox1(+35)TFBS* and *PE Sox1(+35)TFBS+aCGI* (D); or *aCGI* alone (E). For the cells with knock-ins, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*).

4.5 oCGI do not increase the local activation of PEs but rather increase the functional communication of PEs with their target genes

The boosting effect of oCGIs could be attributed to other factors besides its topological role. Based on the notion that CGIs are commonly protected from DNA methylation and display a low nucleosomal density, we reasoned that oCGI could also provide PEs with a permissive environment for TF binding. Therefore, to explore the mechanistic role by which oCGI could potentially facilitate TF binding, we studied the CpG methylation status and the accessibility of the inserted *PE Sox1(+35)* components within *Gata6*-TAD. Bisulfite-qPCR experiments showed that the *PE Sox1(+35)TFBS* insertion is ~40% CpG methylated and the *PE Sox1(+35)TFBS+CGI* is completed hypomethylated in mESC (Figure 4.5.1A). On the other

hand, FAIRE-qPCR experiments in mESC showed that the oCGI only have a mild effect in increasing the chromatin accessibility of their surrounding locus (Figure 4.5.1B). To simultaneously study the chromatin accessibility and CpG methylation levels of the *TFBS*, we performed NOME-seq assays in mESC with the *Sox1(+35)TFBS* or the *Sox1(+35)TFBS+CGI* insertions. This experiment confirmed that oCGIs can protect the nearby TFBS from CpG methylation but do not have major effects in chromatin accessibility (Figure 4.5.1C).

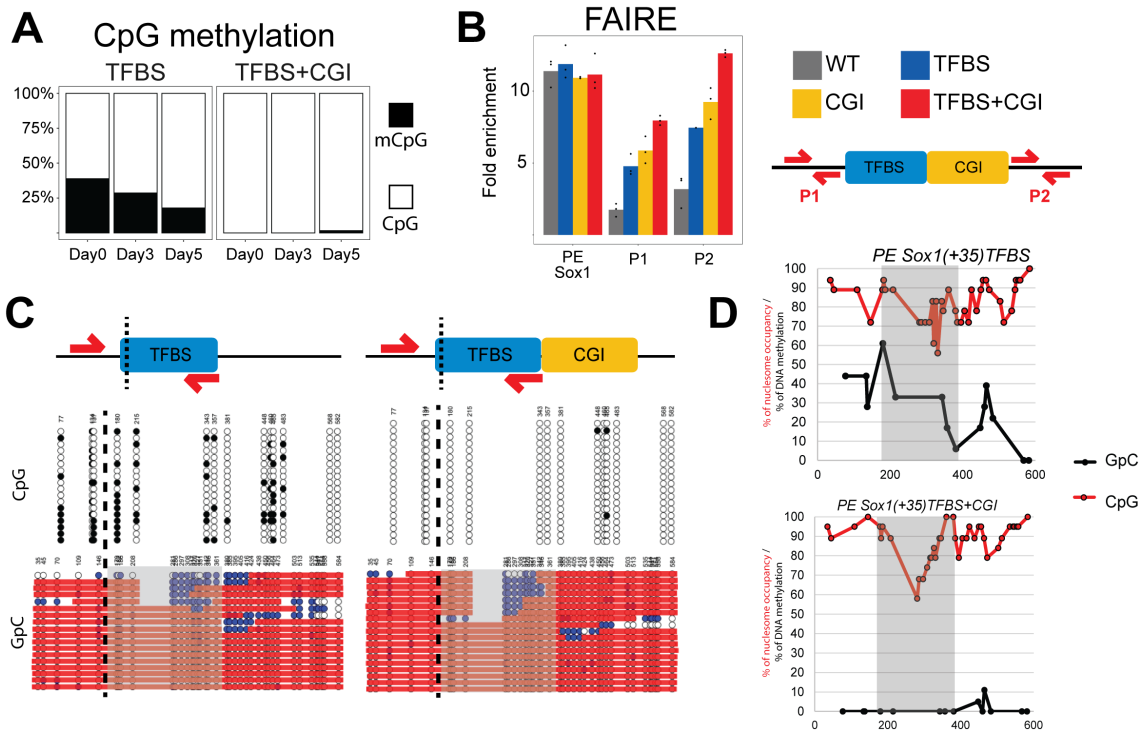


Figure 4.5.1. oCGI protect PEs from CpG methylation but have a mild effect on chromatin accessibility. (A) Bisulfite sequencing analyses during ESCs (Day0) differentiation into AntNPCs (Day5) from cell lines with the *PE Sox1(+35)TFBS* or *PE Sox1(+35)TFBS+CGI* modules inserted in the *Gata6*-TAD. DNA methylation levels were measured using a forward bisulfite primer upstream of the insertion site and a reverse primer inside the TFBS module. **(B)** Chromatin accessibility at the endogenous *PE Sox1(+35)* and the *Gata6*-TAD insertion site (primer pairs P1 and P2) were measured by FAIRE-qPCR in ESCs (left panels) and AntNPC (right panels) that were either WT (gray) or homozygous for the

insertions of the different *PE Sox1(+35)* modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)). FAIRE-qPCR signals were normalized against two negative control regions. The location of the primer pairs P1 and P2 around the *Gata6*-TAD insertion site is represented as red arrows in the diagram shown to the right. **(C)** DNA methylation and nucleosome occupancy at the TFBS module were simultaneously analysed by NOME-PCR in ESC lines with the *PE Sox1(+35)TFBS* (left panel) or *PE Sox1(+35)TFBS+CGI* modules (right panel) inserted in the *Gata6*-TAD. In the upper panels, the black and white circles represent methylated or unmethylated CpG sites, respectively. In the lower panels, the blue or white circles represent accessible or inaccessible GpC sites for the GpC methyltransferase, respectively. Red bars represent regions large enough to accommodate a nucleosome and are consequently considered as inaccessible. The dotted line represents the region where the TFBS sequence starts. The primers used to amplify the TFBS sequences are shown as red arrows in the schematic diagrams, with one of the primers being located within the inserted TFBS and the other one immediately outside. The grey shaded area represents a nucleosome depleted region. **(D)** Scatter plots showing population-averaged nucleosome occupancy (red) and DNA methylation (black) levels within the TFBS sequence in cells with either the *PE Sox1(+35)TFBS* (left panel) or *PE Sox1(+35)TFBS+CGI* (right panel) modules inserted within the *Gata6*-TAD. The grey shaded area represents a nucleosome depleted region.

Importantly, we noticed that the *PE Sox1(+35)TFBS* without the oCGI gets progressively hypomethylated during differentiation into AntNPC (Figure 4.5.1A). This suggest that TFs could still bind to these regions even in a (partially) methylated state and that they might contribute to their demethylation (Grand et al., 2021; Stadler et al., 2011). To test this prediction, we did ChIP-qPCR experiments using antibodies against hallmarks of active enhancers (i.e. p300 and H3K27ac) in the inserted *PE Sox1(+35)* components within *Gata6*-TAD or the *Gria1*-TAD. In the *Gata6-TAD*, *PE Sox1(+35)TFBS* and *PE Sox1(+35)TFBS+CGI* elements showed a similar enrichment for p300 and H3K27ac in AntNPC (Figure 4.5.2A). Notably, in the *Gria1*-TAD, although this gene was not induced by any of the insertions, *PE Sox1(+35)TFBS* and *PE Sox1(+35)TFBS+CGI* elements also gained H3K27ac and p300 binding in AntNPC. Together, these results suggest that the boosting effect of oCGIs do not seem to involve a local effect in the enhancer chromatin.

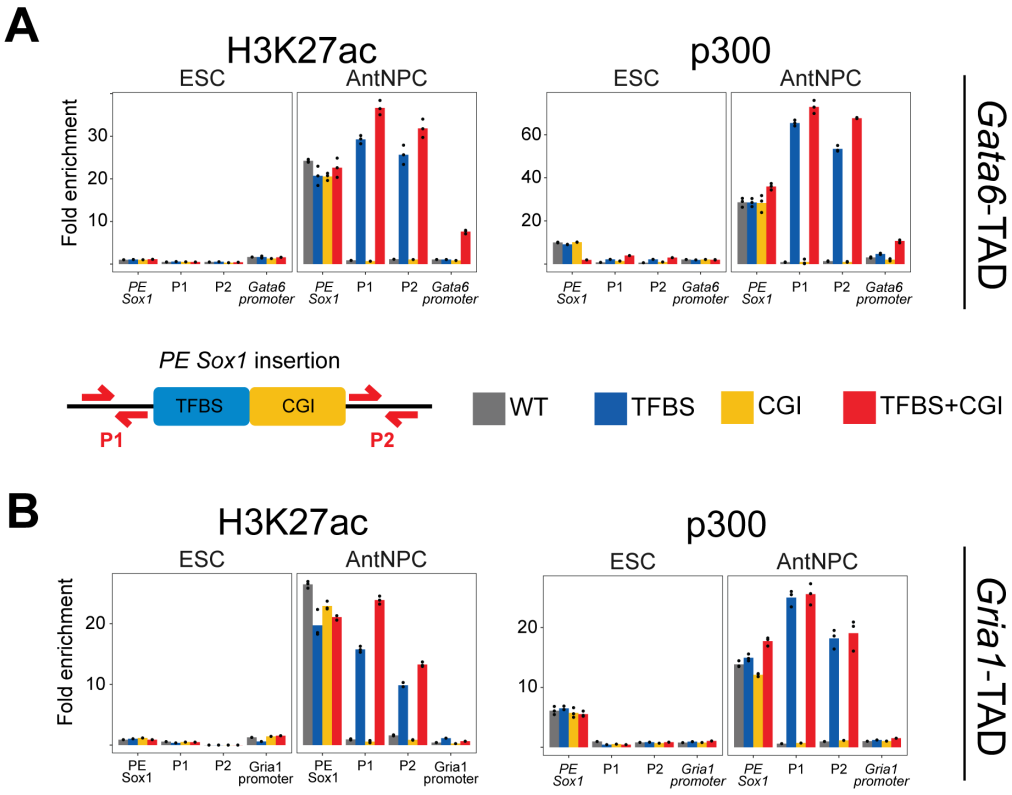


Figure 4.5.2. oCGIs do not influence the local activation of the PE. H3K27ac and p300 levels at the endogenous *PE Sox1(+35)* site, the insertions site (P1 and P2) or the target gene (i.e. *Gata6***(A)** or Gria1**(B)**) were measured by ChIP-qPCR in ESCs (left panels) and AntNPC (right panels) that were either WT (gray) or homozygous for the insertions of the different *PE Sox1(+35)* modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)) at the *Gata6*- **(A)** or the *Gria1*-TAD **(B)**. ChIP-qPCR signals were normalized against two negative control regions. The location of the primers P1 and P2 around the *Gata6*- or the *Gria1*-TAD insertion site are represented as red arrows in the diagram shown in the middle panel.

Another hallmark of enhancer activation is the production eRNAs, which are short bidirectional transcripts generated from the enhancer locus. To investigate whether the oCGIs regulate eRNA production, we did RT-qPCR analysis in ESC and AntNPC using primers flanking the different insertion sites (i.e. *PE Sox1(+35)TFBS, PE Sox1(+35)*CGI and PE *Sox1(+35)TFBS+CGI)* in

Gata6-TAD or *Gria1*-TAD. In contrast to what we observed for the H3K27ac mark, eRNAs levels in the Gata6-TAD were >20-fold higher for the *Sox1(+35)TFBS+CGI* in comparison with the *Sox1(+35)TFBS* insert, while the *Sox1(+35)CGI* did not produce any eRNA (Figure 4.5.3A). In Gria1-TAD, neither of the inserted produced eRNAs (Figure 4.5.3A). Additional ChIP-qPCR experiments in AntNPC revealed that the strong production of eRNAs correlates with a higher enrichment of RNAP2 and the Mediator complex in the *Sox1(+35)TFBS+CGI* in comparison with the *PE Sox1(+35)TFBS* (Figure 4.5.3B). Interestingly, this was only seen in the *Gata6*-TAD insertions, since in *Gria1*-TAD the enrichment of both complexes were very weak (Figure 4.5.3B). Moreover, the *Sox1(+35)CGI* did not recruit neither of these factors (Figure 4.5.3B). These results suggest that oCGIs might increase the functional communication between PEs and genes with CGI-rich promoters.

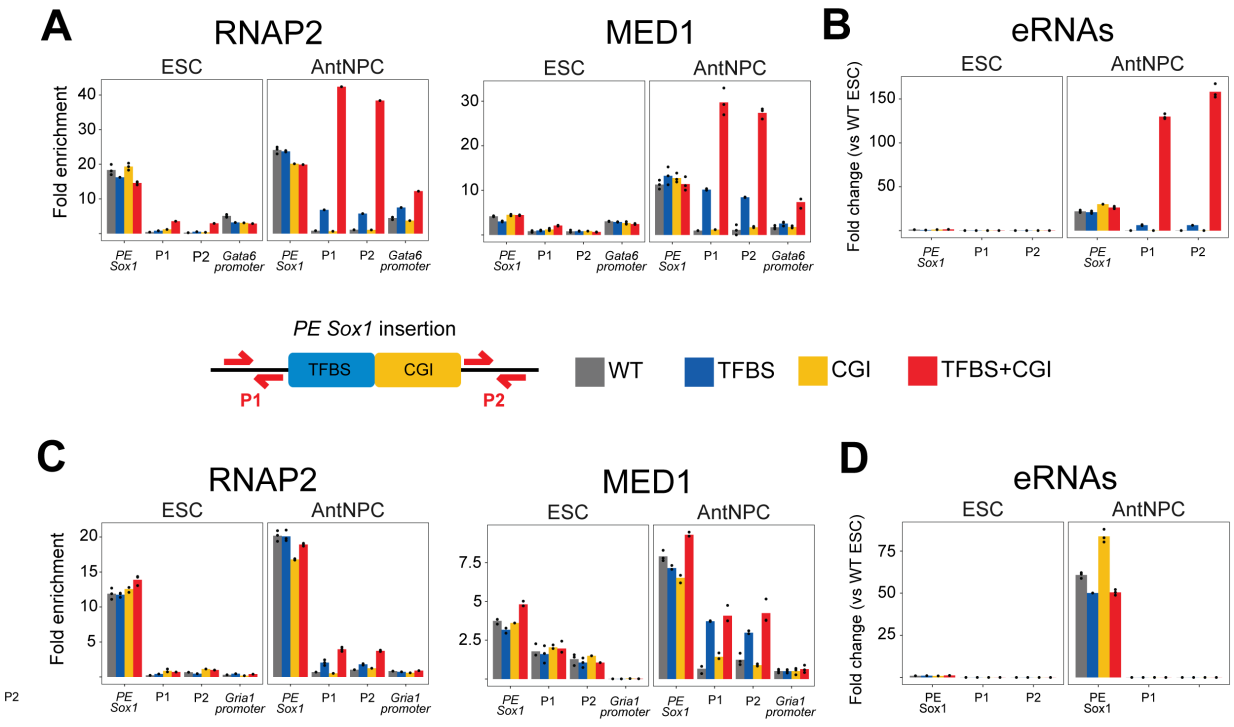


Figure 4.5.3. oCGIs increase the functional communication between PEs and their target genes. (A and C) RNAP2 and MED1 levels at the endogenous *PE Sox1(+35)* site, the insertions site (P1 and P2) or the target gene (i.e. *Gata6***(A)** or Gria1**(C)**) were measured by ChIP-qPCR in ESCs (left panels) and AntNPC (right panels) that were either WT (gray) or homozygous for the insertions of the different *PE Sox1(+35)* modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)) at the *Gata6*- **(A)** or the *Gria1*-TAD **(C)**. ChIP-qPCR signals were normalized against two negative control regions. The location of the primers P1 and P2 around the *Gata6*- or the *Gria1*-TAD insertion site are represented as red arrows in the diagram shown in the middle panel. (**B** and **D**) eRNA levels at the endogenous *PE Sox1(+35)* and the *Gata6*-TAD **(B)** or Gria1-TAD **(D)** insertion sites (primers P1 and P2) were measured by RT-qPCR in ESCs (left panels) and AntNPCs (right panels) that were either WT (gray) or homozygous for the insertions of the different *PE Sox1(+35)* modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)) in the *Gata6*- **(B)** or *Gria*1-TAD **(D)**. The plotted expression values correspond to the average and standard deviation (error bars) from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*) and are presented as fold-changes with respect to WT ESCs.

4.6 oCGI boost PEs cis-activation capacity by increasing the physical communication with their target genes

Altogether, our results suggest that the main regulatory function of oCGIs is to facilitate the establishment of long-range contacts between PEs and CGI-rich genes. This topological configuration might in turn facilitate the establishment of a functional communication between PE and their target gene promoters. In support of this scenario, 4C-seq experiments in AntNPC with the different *PE Sox1(+35)* insertions at the *Gata6*-TAD showed a strong interactions between TFBS+CGI insert and the *Gata6* promoter (Figure 4.6.1). Next, to test the relative contribution of topology to the boosting effect, we inserted PEs modules (i.e. *PE Sox1(+35)TFBS* and *PE Sox1(+35)TFBS+CGI)* in the proximity to *Gria1*-TSS and measure *Gria1* mRNA levels after differentiation into AntNPC (Figure 4.6.2A and B). We reasoned that *Gria1*-TSS is a more appropriate candidate than *Gata6*-TSS, since *Gata6*-TSS already contain CGIs which could mask the contribution of inserted oCGIs. Remarkably, RT-qPCR analysis of the mentioned cell lines demonstrate that both *PE Sox1(+35)TFBS+CGI* and *PE Sox1(+35)TFBS* inserts are able to induce *Gria1* (Figure 4.6.2C). Moreover, the expression of

Gria1 is not boosted by the presence of oCGI. Thus, the contribution of oCGI modules in the cis-activation capacity of PEs might be mainly topological.

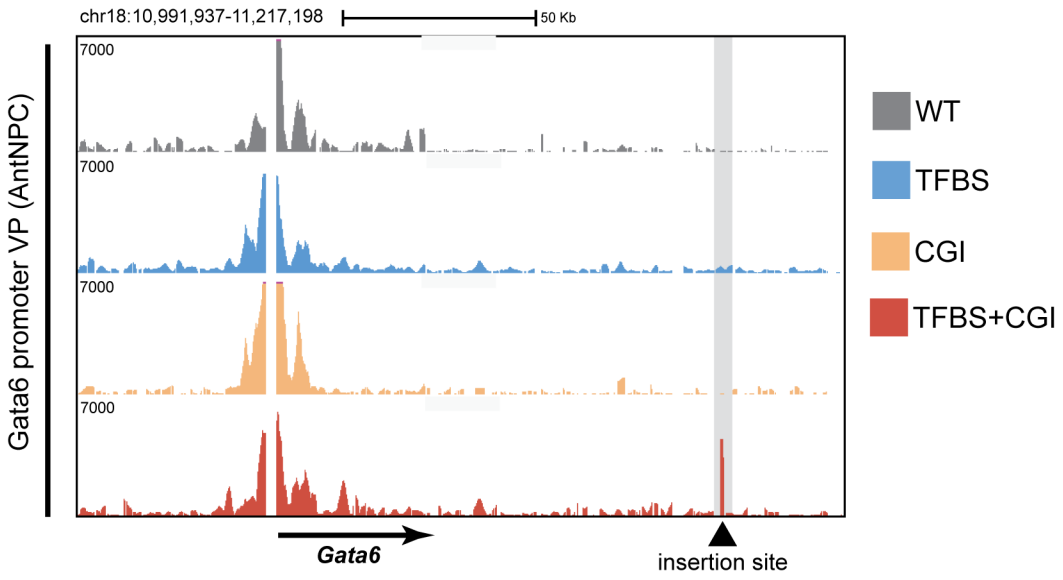


Figure 4.6.1. oCGIs increase the looping between PEs and their target genes in AntNPC. 4C-seq experiments were performed using the Gata6 promoter as a viewpoint in AntNPC that were either WT (black) or homozygous for the insertions of the different PE Sox1(+35) modules (i.e. TFBS (blue), CGI (yellow), TFBS+CGI (red)).

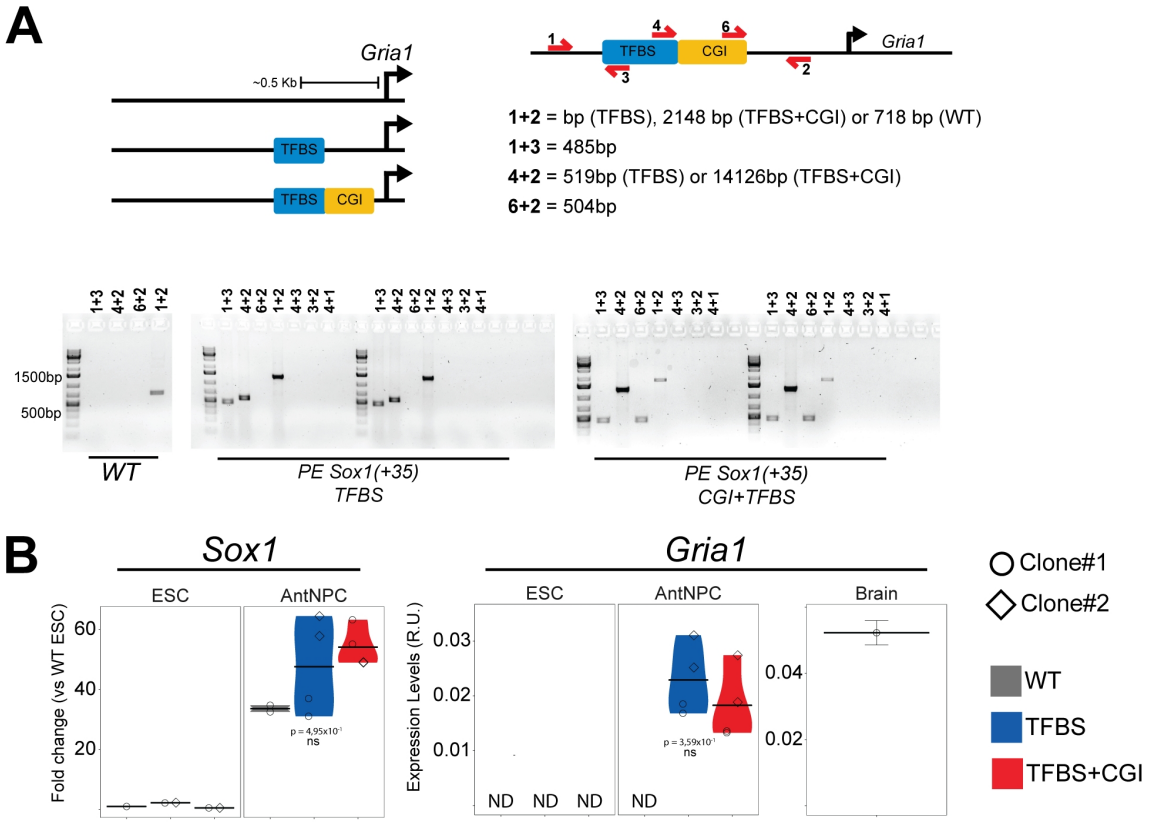


Figure 4.6.2. oCGIs do not have a boosting effect for TSS-proximal enhancers. (A) Insertions of the *PE Sox1(+35)TFBS* or *PE Sox1(+35)TFBS+CGI* at 380bp upstream the *Gria1*-TSS were evaluated by PCR using primers flanking the insertion borders (1+3 and 6+2), amplifying potential duplications (4+3, 3+2 and 4+1) and amplifying a large or small fragment depending on the absence or presence of the insertion (1+2), respectively, were used.. **(B)** RT-qPCR experiments showing the expression levels of *Sox1* and *Gria1* in mESC and AntNPC cell lines that were either WT or have the following insertions at 380bp upstream the TSS: *PE Sox1(+35)TFBS* or *PE Sox1(+35)TFBS+CGI*. For the cells with knock-ins, two different clonal cell lines (circles and diamonds) were studied in each case. For each cell line, two technical replicates of the AntNPC differentiation were performed. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*).

4.7 CGIs and TAD boundaries control gene expression specificity

Our data suggest that intra-TAD homotypic interactions between CGIs located at enhancers and promoters confer a permissive regulatory topology that facilitates gene induction. We hypothesized that the uncovered genetic rules controlling PE-gene compatibility can provide major insights in the understanding of pathomechanisms of human structural variants. To test this prediction, we decided to genetically engineer two loci: the *Six3/Six2* locus and the *Lmx1a/Lrrc52/Mgst3* locus. These loci contain a strong TAD boundary that separates a PE and its target gene from genes having different promoter types (Lenhard et al., 2012) (Figure 4.7.1A and 4.7.2A).

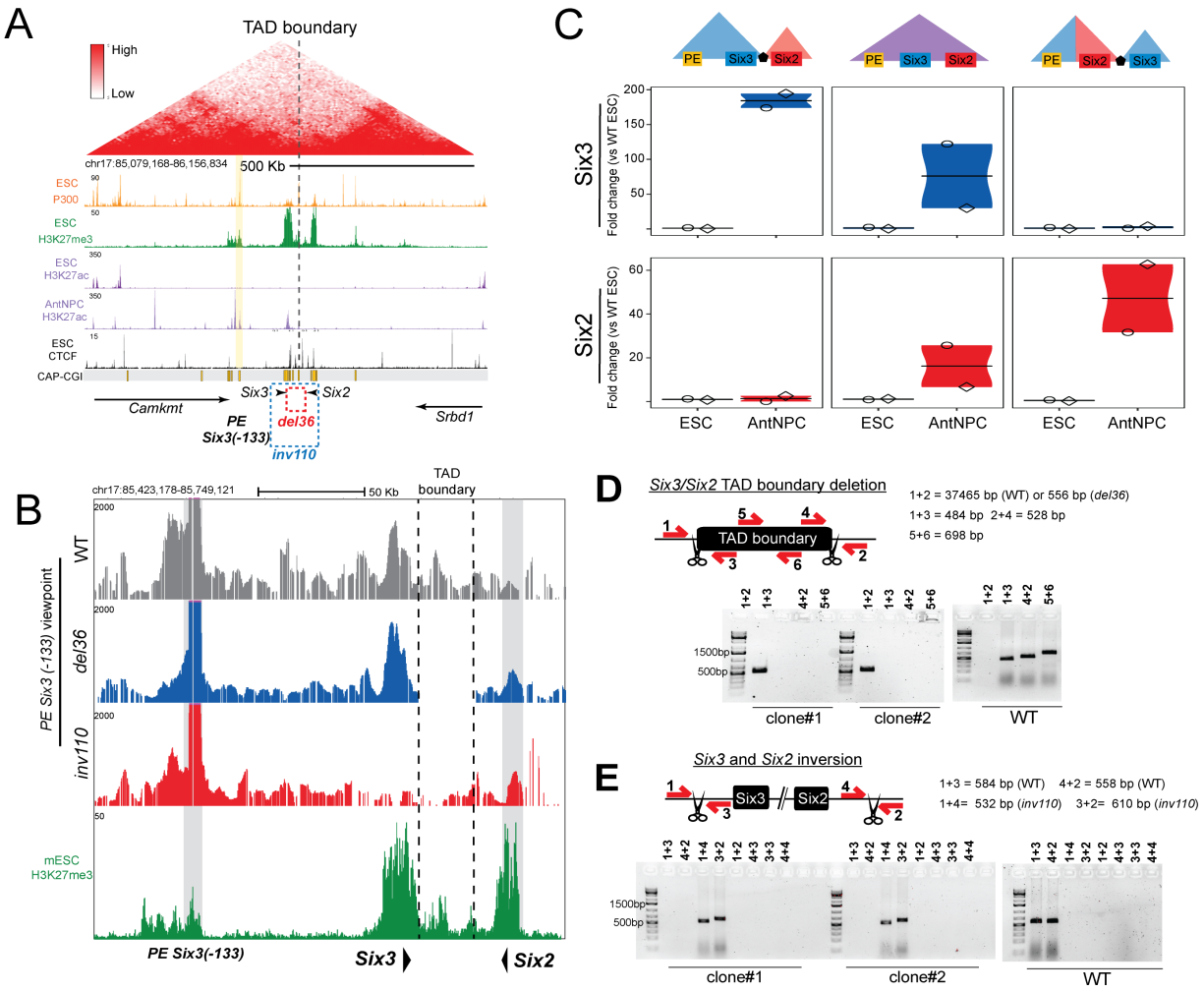


Figure 4.7.1. oCGI and TAD boundaries regulate PEs specific induction of developmental genes. **(A)** The TADs in which *Six3* and *Six2* are located (i.e. *Six3*-TAD and *Six2*-TAD) are shown according to publically available Hi-C data (Bonev et al., 2017). The TAD boundary separating *Six3* and *Six2* is denoted with a dotted line. Below the Hi-C data, several epigenomic and genetic features of the *Six3*-TAD and the *Six2*-TAD are shown. CAP-CGIs (Illingworth et al., 2010) are represented as yellow rectangles. ChIP-seq profiles for the indicated proteins and histone modifications were obtained from (Cruz-Molina et al., 2017). The dotted rectangles indicate the location of the 36 Kb deletion (red) and 110 Kb inversion (blue) engineered in ESCs. (**B)** The expression of *Six3* (blue) and *Six2* (red) was measured by RT-qPCR in ESCs and AntNPCs that were either WT (left panel), homozygous for the 36 Kb deletion (del36; middle panel) or homozygous for the 110 Kb inversion (inv110; right panel). For each of the engineered structural variants, two different clonal cell lines were generated and independently differentiated into AntNPCs. The plotted expression values for each clone correspond to the average from three RT-qPCR technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*) and represented as fold-changes with respect to WT ESCs. **(C)** 4C-seq experiments were performed using the *PE Six3(-133)* as viewpoint in ESCs that were either WT (grey), homozygous for the 36 Kb deletion (del36; blue) or homozygous for the 110 Kb inversion (inv110; red). ChIP-seq profiles of H3K27me3 are shown in green (Cruz-Molina et al., 2017). **(D)** For the identification of ESC lines with the *Six3/Six2* TAD boundary deletion, primer pairs flanking the deleted region (1+3 and 4+2), amplifying the deleted fragment (5+6) and amplifying a large or small fragment depending on the absence or presence of the deletion (1+2), respectively, were used. The PCR results obtained for two ESC clonal lines with 36Kb homozygous deletions (*del36*) are shown. **(E)** For the identification of ESC lines with the *Six3/Six2* inversion, primer pairs flanking the inverted region (1+3, 4+2, 1+4 and 3+2) and amplifying potential duplications (4+3, 3+3 and 4+4) were used. The PCR results obtained for two ESC clonal lines with 110 Kb homozygous inversions (*inv110*) are shown.

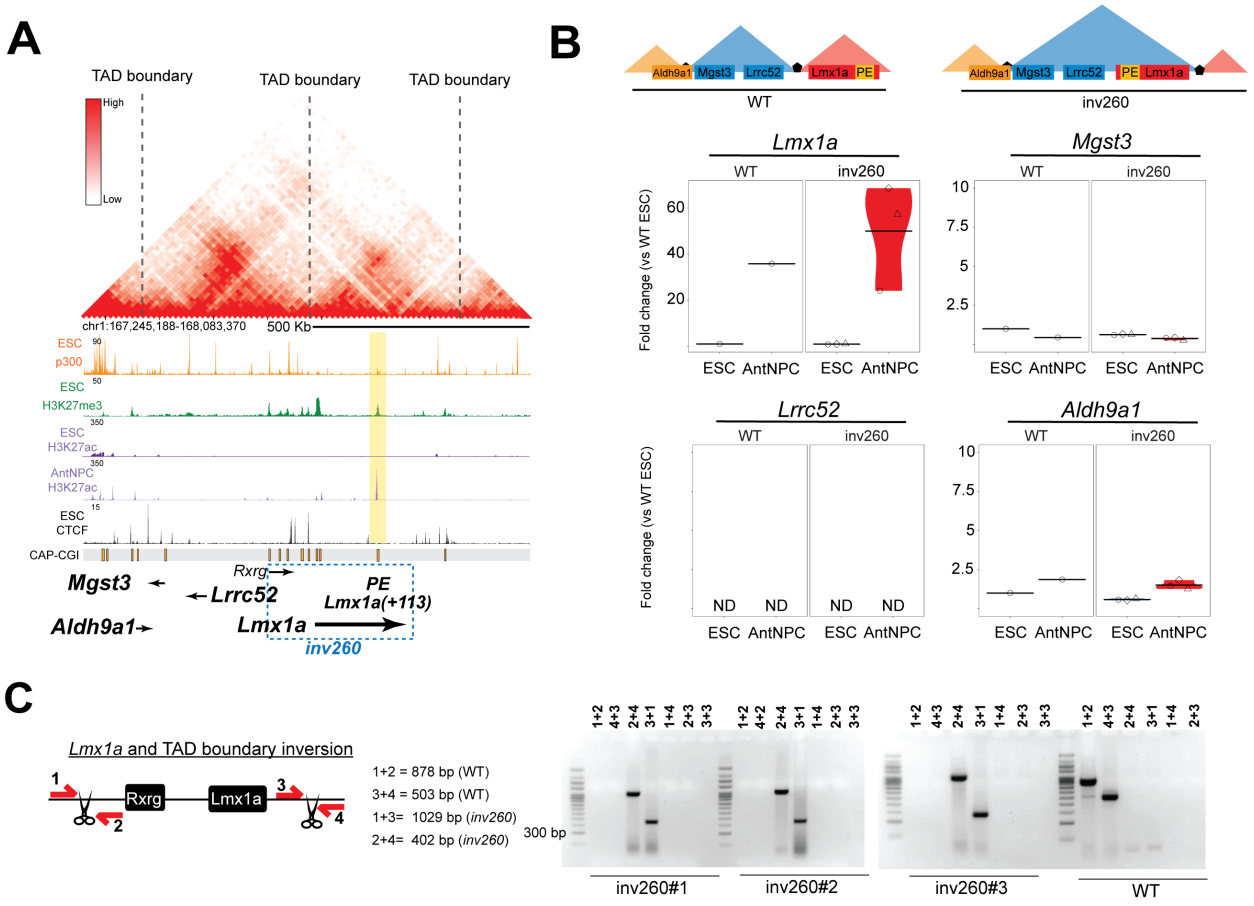


Figure 4.7.2. Housekeeping and tissue-specific genes are not responsive to PE cis-activation. (A) The TADs in which L*mx1a*, *Lrrc52* and *Mgst3* are located are shown according to publically available Hi-C data (Bonev et al., 2017); the TAD boundaries are denoted with a dotted line. Below the Hi-C data, several epigenomic and genetic features of the corresponding TADs are shown. CAP-CGIs (Illingworth et al., 2010) are represented as yellow rectangles. ChIP-seq profiles for the indicated proteins and histone modifications were obtained from (Cruz-Molina et al., 2017). The dotted rectangle indicates the location of the 260 Kb inversion (blue; inv260) engineered in ESCs. **(B)** The expression of *Lmx1a, Mgst3*, *Lrrc52* and *Aldh9a1* was measured by RT-qPCR in ESCs and AntNPCs that were either WT (left panel) or homozygous for the inv260 (right panels). For the inv260, two different clonal cell lines were generated and independently differentiated into AntNPCs. The plotted expression values for each clone correspond to the average from three technical replicates. Expression values were normalized to two housekeeping genes (*Eef1a* and *Hprt*) and represented as fold changes with respect to WT ESCs **(C)** For the identification of ESC lines with the *Lmx1a*-TAD boundary inversion, primer pairs flanking the inverted region (1+3, 4+2, 1+4 and 3+2) and amplifying potential deletions (1+4) or duplications (4+3, 3+3 and 4+4) were used. The PCR results obtained for three ESC clonal lines with 260 Kb homozygous inversions (*inv260*) are shown.

RESULTS

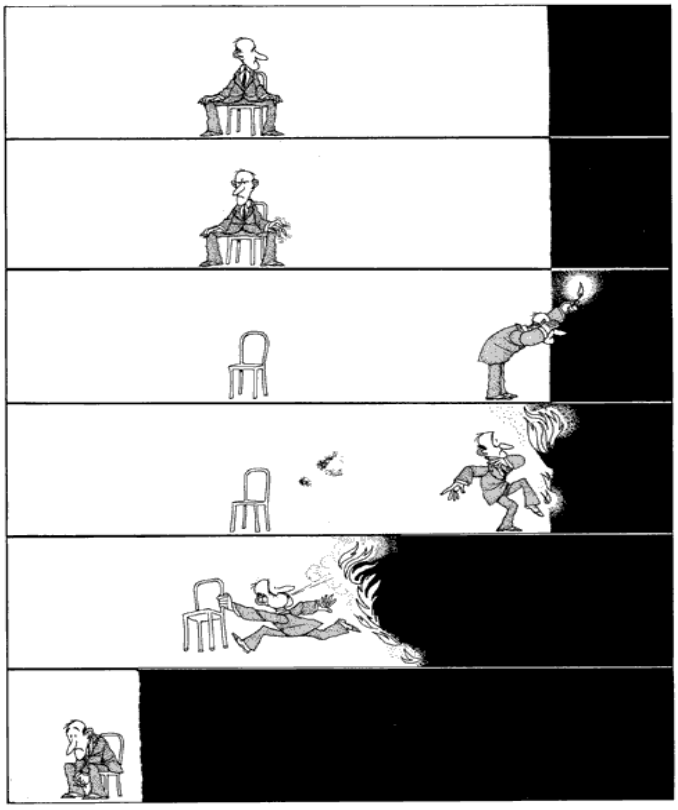
We first focused on the *Six3/Six2* locus, which was selected due to several reasons: (i) *Six3* and *Six2* are next to each other, yet they are contained within two neighboring TADs separated by a conserved TAD boundary; (ii) *Six3* and *Six2* display mutually exclusive expression patterns during embryogenesis (e.g. *Six3* in brain; *Six2* in facial mesenchyme); (iii) the *Six3* TAD contains a PE (i.e. *PE Six3(-133)*) that controls the induction of *Six3* in AntNPC without any effects on *Six2*; (iv) in mESC, the *PE Six3(-133*) strongly interacts with *Six3* but not with *Six2* although both genes contain multiple pCGI. Taking all this information into account, we generated mESC with two different homozygous genomic rearrangements: (i) a 36 Kb deletion spanning the *Six3/Six2* TAD boundary (*del36*) and (ii) a 110 Kb inversion that places *Six3* within the *Six2* TAD and *vice versa* (*inv110*) (Figure 4.7.1A and B). Interestingly, upon differentiation into AntNPC, *Six2* expression strongly increased in *del36* and *inv110* cells; while the expression of *Six3* expression was dramatically reduced in *inv110* cells and mildly affected in *del36* cells (Figure 4.7.1C). Furthermore, and in agreement with the previous gene expression changes, 4C-seq experiments in WT, *del36* or *inv110* ESC showed that both the deletion of the *Six3/Six2* boundary and the inversion of *Six3* and *Six2* resulted in increased interactions between *Six2* and the *PE Six3(-133)* (Figure 4.7.1D). In addition, *Six3* decreased its interaction with *PE Six3(-133)* in *inv110* ESC.

Next, we focused on the *Lmx1a/Lrrc52/Mgst3* locus, which was selected due to the following reasons (Figure 4.7.2A): (i) *Lmx1a* and *Lrrc52/Mgst3* are located in neighboring TADs separated by a strong TAD boundary; (ii) the three genes have different types of promoters (Lenhard et al., 2012) (i.e. *Lmx1a* has a type III promoter, which contains a large CGI cluster; *Lrrc52* has a type I promoter, which does not contain any CGIs; and *Mgst3* has a type II promoter that contains a single and short CGI centered on its TSS); (iii) the *Lmx1a*-TAD contains a PE (i.e. *PE Lmx1a(+113)*) that becomes strongly activated in AntNPC and that presumably contributes to *Lmx1a* induction in these cells. Considering this information, we generated two mESC homozygous clones with a 260 Kb inversion that places the *Lmx1a* gene and the *PE Lmx1a(+113)* within the *Lrrc52/Mgst3*-TAD (*inv260*) (Figure 4.7.2A and B). In this case, differentiation of *inv260* into AntNPC did not result in ectopic gene induction (Figure 4.7.2C).

That is, neither *Lrrc52* nor *Mgst3* were induced by *PE Lmx1a(+113)*, which suggest that tissue specific genes and housekeeping genes are not responsive to distal PEs.

Overall, these results indicate that PEs can specifically execute their regulatory functions due to the combined effects of TAD boundaries, which provide insulation, and the long-range homotypic interactions between oCGIs and pCGIs, which confer enhancer responsiveness (Figure 4.7.3).

5- DISCUSSION AND FUTURE WORK



by Quino (1983).

Understanding how genome 3D structures organizes regulatory domains and how enhancers specifically refine this interaction towards their target genes is still an ongoing challenge. Addressing this questions would require novel approaches than the classical enhancer-reporter assays, in which enhancers and promoters are placed close to each other. Here, we satisfy this requirement by using a genetic knock-in approach to identify genetic elements that dictate the topological and regulatory compatibility between PEs and their target genes.

DISCUSSION AND FUTURE WORK

The current model of enhancer-gene regulation assumes that enhancers can regulate genes regardless of their distance or sequence composition as long as there are located within the same TAD (Galupa & Crocker, 2020; Ibrahim & Mundlos, 2020; Kragesteen et al., 2018; Lupiáñez et al., 2016; Nora et al., 2012). That is, TAD boundaries are the main regulatory element that prevents ectopic communication between enhancers and non-target genes. However, this model has been challenged by several recent studies, which demonstrate that placing enhancers and genes into the same TAD is not always sufficient for productive transcription (Ghavi-Helm et al., 2019; Laugsch et al., 2019; Rao et al., 2017a). This suggests that there must be additional genetic and biochemical factors that could contribute to regulate the compatibility between genes and enhancers. In agreement with this scenario, massive-parallel report assays in Drosophila showed that core promoter motifs separate genes to independently be responsive to either housekeeping or developmental enhancers (Arnold et al., 2016; Zabidi et al., 2015). Here, we showed that homotypic genetic elements, located both at distal enhancers and promoters, can also affect the responsiveness of target genes. Namely, oCGIs and pCGI can engage into long-range interactions that facilitate the topological and functional communication between enhancers and genes. This configuration in turn make CGI-rich genes particularly responsive to PEs, in opposition to CGI-poor genes.

The remaining question is whether CGI elements are the solely responsible for regulating the compatibility between PEs and their target genes. Our attempt to make a tissue-specific gene (i.e. Gria1) responsive to PEs by inserting a single pCGI was not sufficient (data not shown). Thus, additional sequences should be inserted into the promoter of this gene to test which combination of genetic elements could make a tissue-specific gene responsive to a distal enhancer. It is important to note that PEs can activate a tissue-specific gene when placed in close proximity. Therefore, the missing promoter genetic elements necessary for enhancer responsiveness should probably regulate enhancer-promoter physical communication. We argue that a promising candidate for making *Gria1* responsive to the distal PE would be the insertion of more than one pCGI, since developmental genes usually contain clusters of pCGI (Lenhard et al., 2012).

Moreover, by doing an inversion that places a PE into the TAD of a housekeeping gene, we demonstrate that promoters from housekeeping genes are also not responsive to PEs. The fact that housekeeping genes do not respond to PEs, but usually contain a CGI in its promoter (Lenhard et al., 2012) could be explain by different reasons: (i) housekeeping genes contain core promoters that are not compatible to developmental enhancers (Arnold et al., 2016; Haberle et al., 2019; Zabidi et al., 2015); (ii) housekeeping genes are actively transcribed across different cellular states, including ESC, and therefore might not be able to engage into PcG-dependent interactions with PEs; (iii) the presence of a single pCGI at housekeeping promoters (Lenhard et al., 2012) might not be sufficient to respond to PEs; (iv) the housekeeping genes are already expressed at high levels that enhancers can not increase any further. It would be exciting to test these hypothesis by engineering housekeeping gene promoters and then test their ability to respond to distal PEs.

From a mechanistic point of view, our data suggest that the boosting capacity of oCGI might be mainly attributed to its topological functions. That is, physical interactions between oCGI and pCGI located at PE target genes. In ESC, these interactions are mediated by long-range PcG-PcG interactions (Cruz-Molina et al., 2017; Entrevan et al., 2016; Pachano, Crispatzu, & Rada-iglesias, 2019), likely via the polymerization capacity of the PRC1 subunit PHC1 (Isono et al., 2013). Therefore, the lack of responsiveness of tissue-specific genes to PEs could be mechanistically explain by the fact that they are not enriched in PRC1 and consequently they do not physically interact with PEs. Interestingly, insertion of a single pCGI at *Gria1* promoter did not recruit PRC1 levels as the same extend as usually found in developmental genes. We hypothesize that the insertion of an additional pCGI could result in greater PRC1 levels that could mediate a more stable interactions between the PE and the tissue-specific gene. However, it is important to note the fact that an oCGI element was not sufficient to recruit PRC1 when inserted alone. The recruitment of PRC1 by the oCGI required the presence of a nearby TFBS. Therefore, we can not dismiss the possibility that additional genetic elements apart from CGIs would be require to enrich tissue-specific promoters in PRC1.

Independently from the facilitator role of PcG (Cruz-Molina et al., 2017), the boosting capacity of oCGI could also attributed to transcriptional activators that specifically recognize CGIs. In agreement, pCGIs can also be bound by CxxC-domain containing protein involved in transcriptional activation, like MLL1/2, CFP1 and KDM2A (Farcas et al., 2012; Wu et al., 2013). It would be interesting to asses for the presence of these factors in active PE and test whether they also contribute to the boosting effect of oCGIs.

Our data suggest that the boosting capacity of oCGIs might not be attributed to an increase in TF binding, but rather in facilitating the functional communication between TFs and target genes. By engaging into homotypic interactions with pCGI, oCGI could keep PEs and their target genes in close proximity as the time the TFs binds their cognate sequences at the PE. This would ensure a precise and uniform induction of target genes upon differentiation. Accordingly, we showed that oCGI increase the binding of Mediator and RNAP2 upon PE activation. Since these complexes have the ability to form phase-separated condensates (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018), oCGIs might facilitate the inclusion of PEs and their target genes into phase-separated condensates by bringing both elements into close proximity by the time the activation cues appear. It would be important to test whether oCGIs also contribute to enhancer function after the cis-activation of the target gene. That is, oCGIs could be necessary to maintain gene expression.

In conclusion, we propose a model whereby the specific induction of developmental genes is achieved through the combination of CGI-mediated long-range chromatin interactions and the insulation provided by TAD boundaries (Figure 5.1).We anticipate that our findings can help to better understand pathomechanisms of human congenital diseases caused by structural variants. For example, a heterogeneous inversion found in a patient with Branchio-oculo-facial syndrome (BOFS) (Laugsch et al., 2019). This inversion places one of the *TFAP2A* alleles into a novel TAD and impairs its normal expression in neural crest cells due to the physical disconnection from its cognate neural crest enhancers (Figure 5.2A). The *TFAP2A* gene is an archetypical developmental gene with a promoter containing a large CGI cluster and marked with H3K27me3 in ESC. Moreover, some of its cognate neural crest enhancers are associated with oCGI and are

also marked with H3K27me3. Interestingly, this inversion also places novel genes whose promoters contained a short CGI centered on their TSS (a typical feature of housekeeping genes(Lenhard et al., 2012)) in proximity of the *TFAP2A* neural crest enhancers within a shuffled topological domain. In agreement with our findings, none of these genes is responsive to the *TFAP2A* neural crest enhancers (*i.e.* no enhancer adoption in neural crest cells). Another example corresponds to a deletion found in families with brachydactyly (Lupiáñez et al., 2015). This deletion includes *EPHA4, a* gene highly expressed in the developing limb, along with a TAD boundary located between *EPHA4*-TAD and *PAX3*-TAD (Figure 5.2B). As a result, enhancers active in the developing limb and that control *EPHA4* expression in this tissue, establish ectopic interaction with *PAX3* (i.e. enhancer adoption) and strongly induce its expression in the limb, a tissue in which this developmental gene is normally inactive. Notably, *PAX3* is a gene whose promoter is marked with H3K27me3 in ESC and that contains a large CGI cluster, while one of the major *EPHA4* enhancers is associated with an oCGI and is also mark with H3K27me3. Here again the high responsiveness of *PAX3* to the *EPHA4* enhancers is in agreement with our findings.

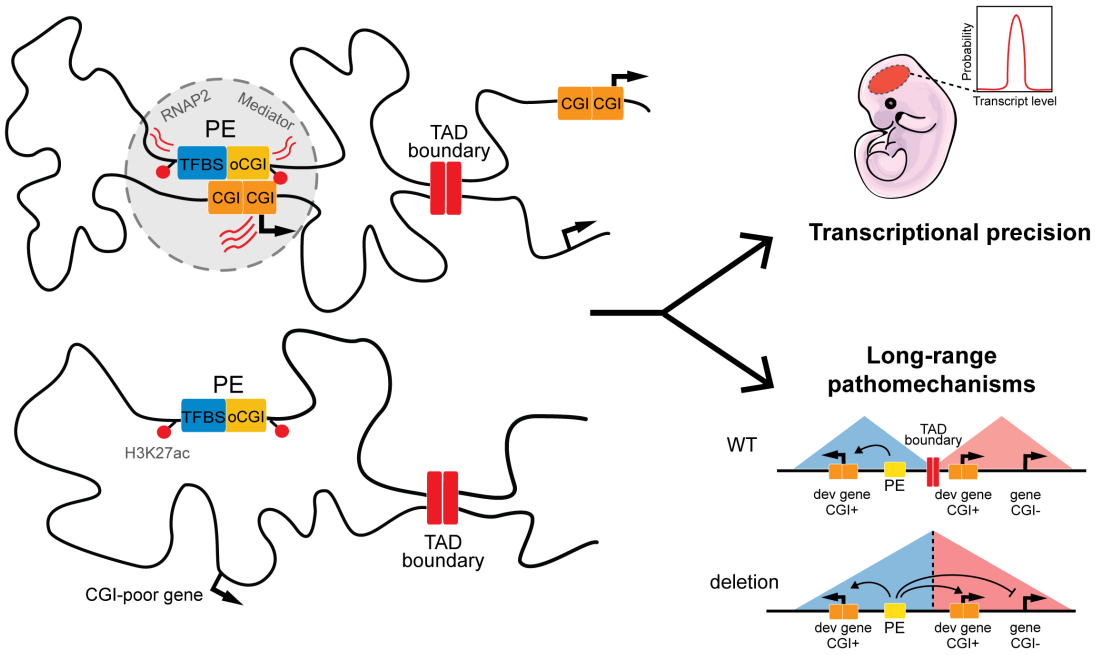


Figure 5.1. Proposed model for oCGI function in PEs. oCGIs increase the physical and functional communication of PEs with their target genes due to homotypic interactions between oCGI and proximal pCGI clusters. Consequently, developmental genes with CGI-rich promoters will be preferentially responsive to PEs, in comparison with tissue-specific or housekeeping genes that lack pCGI clusters. In addition, TAD boundaries add another layer in the regulation of PE specificity by restricting interactions between oCGI and pCGI from neighboring TADs.

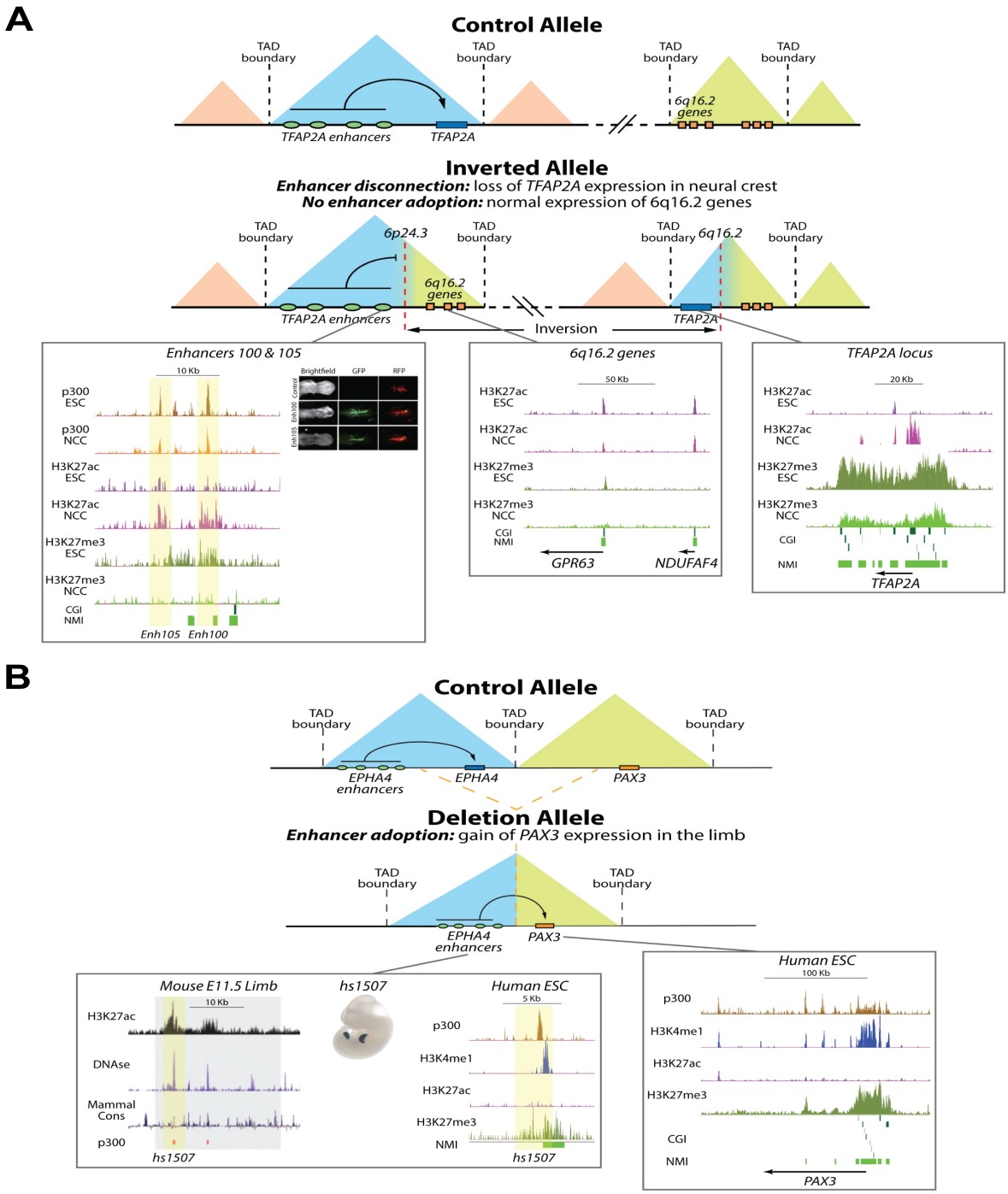

A
Control Allele
TAD boundary
TAD boundary
TAD boundary
TAD boundary
6q16.2 genes
TFAP2A enhancers
TFAP2A
Inverted Allele
Enhancer disconnection: loss of TFAP2A expression in neural crest
No enhancer adoption: normal expression of 6q16.2 genes
6p24.3
6q16.2
6q16.2 genes
Inversion
Enhancers 100 & 105
10 Kb
Brightfield
GFP
RFP
p300 ESC
p300 NCC
H3K27ac ESC
H3K27ac NCC
H3K27me3 ESC
H3K27me3 NCC
CGI
NMI
Enh105
Enh100
6q16.2 genes
50 Kb
GPR63
NDUFAF4
TFAP2A locus
20 Kb
TFAP2A
B
Control Allele
EPHA4 enhancers
EPHA4
PAX3
Deletion Allele
Enhancer adoption: gain of PAX3 expression in the limb
Mouse E11.5 Limb
hs1507
Human ESC
5 Kb
DNAse
Mammal Cons
100 Kb
H3K4me1

Figure 5.2. Two examples of human congenital diseases caused by structural variants that disrupt developmental loci with PE-associated oCGIs. (A) Upper panel: diagram of a heterozygous inversion found in a patient with Branchio-oculo-facial syndrome (BOFS) involving the *TFAP2A* locus (Laugsch et al., 2019). Lower panel: epigenomic and genetic features of *TFAP2A* neural crest cognate enhancers (left), 6q16.2 genes (middle) and *TFAP2A* (right). In the lower left panel, enhancer reporter assays for two representative *TFAP2A* enhancers are shown in stage HH14-16 chick embryos. Computational and biochemically identified CGI (CGI and NMIs (R. S. Illingworth et al., 2010), respectively) are represented as green rectangles. The CGIs were defined in the UCSC browser according to the following criteria: GC content > 50%; Length > 200 bp; CpG Observed to expected ratio > 0.6. The inversion places one of the *TFAP2A* alleles into a novel TAD and impairs its normal expression in neural crest cells due to the physical disconnection from its enhancers. *TFAP2A* is a developmental gene with a promoter containing a large CGI cluster and marked with a broad H3K27me3 domain in ESCs. Some of the *TFAP2A* neural crest enhancers are associated with oCGI/NMI and are also marked with H3K27me3 in ESCs. Moreover, this inversion also places novel genes originally found within the 6q16.2 locus in proximity of the *TFAP2A* neural crest enhancers within a shuffled topological domain. The promoters of these 6q16.2 genes (i.e *GPR63* and *NDUFAF4*) contained only a short CGI centered on their TSS. In agreement with our findings, none of the 6q16.2 genes was found to be responsive to the *TFAP2A* neural crest enhancers (*i.e.* no enhancer adoption in neural crest cells). **(B)** Upper panel: diagram of a deletion found in families with brachydactyly involving a TAD boundary located between the *EPHA4* and the *PAX3* loci (Lupiáñez et al., 2015). Lower panel: epigenomic and genetic features of *the Epha4* cognate enhancers in the mouse E11.5 limb (left) and in human ESCs (right). In addition, a representative E11.5 transgenic embryo with limb-specific expression of a reporter gene driven by the *hs1507* element is shown in the middle. CGIs and NMIs are defined as in (A). This deletion includes *EPHA4, a* gene highly expressed in the developing limb, along with a TAD boundary separating the *EPHA4*-TAD and *PAX3*-TAD. As a result, enhancers that control *EPHA4* expression in the limb establish ectopic interactions with *PAX3* (i.e. enhancer adoption) and strongly induce its expression in the limb, a tissue in which this developmental gene is normally inactive. Notably, *PAX3* is a gene whose promoter is marked with H3K27me3 in ESCs and that contains a large CGI cluster, while one of the major *EPHA4* enhancers (*hs1507*) is associated with an oCGI/NMI and is marked with H3K27me3 in ESCs. Here again the high responsiveness of *PAX3* to the *EPHA4* enhancers is in agreement with our findings.

6- REFERENCES

Abdennur, N., & Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, *36*(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540

Akasaka, T., van Lohuizen, M., van der Lugt, N., Mizutani-Koseki, Y., Kanno, M., Taniguchi, M., Vidal, M., Alkema, M., Berns, A., & Koseki, H. (2001). Mice doubly deficient for the polycomb group genes Mel18 and Bmi1 reveal synergy and requirement for maintenance but not initiation of Hox gene expression. *Development*, *128*(9), 1587–1597. https://doi.org/10.1242/dev.128.9.1587

Alexander, J. M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S., & Weiner, O. D. (2019). Live-cell imaging reveals enhancer-dependent sox2 transcription in the absence of enhancer proximity. *ELife*, *8*, 1–42. https://doi.org/10.7554/eLife.41769

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., & Stark, A. (2016). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature Biotechnology*, *35*(2), 136–144. https://doi.org/10.1038/nbt.3739

Bajic, V. B., Sin, L. T., Christoffels, A., Schönbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D. A., Carninci, P., & Hayashizaki, Y. (2006). Mice and men: Their promoter properties. *PLoS Genetics*, *2*(4), 614–626. https://doi.org/10.1371/journal.pgen.0020054

Bell, J. S. K., & Vertino, P. M. (2017). Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics*, *12*(0), 449–464. https://doi.org/10.1080/15592294.2017.1297910

Benabdallah, N. S., Williamson, I., Illingworth, R. S., Grimes, G. R., Therizols, P., & Bickmore, W. A. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular Cell*, 1–12. https://doi.org/10.1016/j.molcel.2019.07.038

Bird, A. P., & Wolffe, A. P. (1999). Methylation-induced repression-belts, braces, and chromatin. *Cell*, *99*(5), 451–454. https://doi.org/10.1016/S0092-8674(00)81532-9

Bird, A., Taggart, M., Frommer, M., Miller, O. J., & Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, *40*(1), 91–99. https://doi.org/10.1016/0092-8674(85)90312-5

REFERENCES

Blackledge, N. P., Farcas, A. M., Kondo, T., King, H. W., McGouran, J. F., Hanssen, L. L. P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., Ishikura, T., Long, H. K., Sheahan, T. W., Brockdorff, N., Kessler, B. M., Koseki, H., & Klose, R. J. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell*, *157*(6), 1445–1459. https://doi.org/10.1016/j.cell.2014.05.004

Blackledge, N. P., & Klose, R. J. (2021). The molecular principles of gene regulation by Polycomb repressive complexes. *Nature Reviews Molecular Cell Biology*, *0123456789*. https://doi.org/10.1038/s41580-021-00398-y

Blackledge, N. P., Long, H. K., Zhou, J. C., Kriaucionis, S., Patient, R., & Klose, R. J. (2012). Bio-CAP: A versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Research*, *40*(4). https://doi.org/10.1093/nar/gkr1207

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B. J., Afeyan, L. K., Guo, Y. E., Rimel, J. K., Fant, C. B., Schuijers, J., Lee, T. I., Taatjes, D. J., & Young, R. A. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, *175*(7), 1842-1855.e16. https://doi.org/10.1016/j.cell.2018.10.042

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–679. https://doi.org/10.1038/nrg.2016.112

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., & Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, *171*(3), 557–572. https://doi.org/10.1016/j.cell.2017.09.043

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., & Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, *44*(2), 148–156. https://doi.org/10.1038/ng.1064

Bothma, J. P., Garcia, H. G., Ng, S., Perry, M. W., Gregor, T., & Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *ELife*, *4*, 1–14. https://doi.org/10.7554/elife.07956

Boulard, M., Edwards, J. R., & Bestor, T. H. (2015). FBXL10 protects Polycomb-bound genes from hypermethylation. *Nature Genetics*, *47*(5), 479–485. https://doi.org/10.1038/ng.3272

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). 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. https://doi.org/10.1038/nmeth.2688

REFERENCES

Butler, J. E. F., & Kadonaga, J. T. (2001). *Enhancer – promoter specificity mediated by DPE or TATA core promoter motifs*. 2515–2519. https://doi.org/10.1101/gad.924301.protein

Cho, W. K., Spille, J. H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, *361*(6400), 412–415. https://doi.org/10.1126/science.aar4199

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J. H., Skalnik, D., & Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development*, *26*(15), 1714–1728. https://doi.org/10.1101/gad.194209.112

Cohen, I., Zhao, D., Bar, C., Valdes, V. J., Dauber-Decker, K. L., Nguyen, M. B., Nakayama, M., Rendl, M., Bickmore, W. A., Koseki, H., Zheng, D., & Ezhkova, E. (2018). PRC1 Fine-tunes Gene Repression and Activation to Safeguard Skin Development and Stem Cell Specification. *Cell Stem Cell*, *22*(5), 726-739.e7. https://doi.org/10.1016/j.stem.2018.04.005

Cohen, I., Zhao, D., Menon, G., Nakayama, M., Koseki, H., Zheng, D., & Ezhkova, E. (2019). PRC1 preserves epidermal tissue integrity independently of PRC2. *Genes and Development*, *33*(1–2), 55–60. https://doi.org/10.1101/gad.319939.118

Cooper, D., Taggart, M. H., & Bird. (1983). Unmethylated domains in vertebrate DNA. *Nucleic Acids Research*, *11*(3), 647–658.

Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N. P., DeMarco, V., Elderkin, S., Koseki, H., Klose, R., Heger, A., & Brockdorff, N. (2014). Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. *Cell Reports*, *7*(5), 1456–1470. https://doi.org/10.1016/j.celrep.2014.04.012

Creppe, C., Palau, A., Malinverni, R., Valero, V., & Buschbeck, M. (2014). A Cbx8-Containing Polycomb Complex Facilitates the Transition to Gene Activation during ES Cell Differentiation. *PLoS Genetics*, *10*(12). https://doi.org/10.1371/journal.pgen.1004851

Crispatzu, G., Rehimi, R., Pachano, T., Bleckwehl, T., Cruz-Molina, S., Xiao, C., Mahabir, E., Bazzi, H., & Rada-iglesias, A. (2021). The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nature Communications*, *12*(4344), 1–17. https://doi.org/10.1038/s41467-021-24641-4

Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W. F. J., Grosveld, F., Frommolt, P., Bazzi, H., & Rada-Iglesias, A. (2017). PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell*, *20*, 1–17. https://doi.org/10.1016/j.stem.2017.02.004

REFERENCES

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, *295*, 1306–1311. https://doi.org/10.1007/978-1-0716-0664-3_1

Denker, A., & De Laat, W. (2016). The second decade of 3C technologies: Detailed insights into nuclear organization. *Genes and Development*, *30*(12), 1357–1382. https://doi.org/10.1101/gad.281964.116

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R. Z., Ragozin, S., & Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *Journal of Biological Chemistry*, *285*(34), 26114–26120. https://doi.org/10.1074/jbc.M109.089433

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., Weintraub, A. S., Schuijers, J., Lee, T. I., Zhao, K., & Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, *159*(2), 374–387. https://doi.org/10.1016/j.cell.2014.09.030

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, *539*(7629), 452–455. https://doi.org/10.1038/nature20149

Entrevan, M., Schuettengruber, B., & Cavalli, G. (2016). Regulation of Genome Architecture and Function by Polycomb Proteins. *Trends in Cell Biology*, *26*(7), 511–525. https://doi.org/10.1016/j.tcb.2016.04.009

Farcas, A. M., Blackledge, N. P., Sudbery, I., Long, H. K., McGouran, J. F., Rose, N. R., Lee, S., Sims, D., Cerase, A., Sheahan, T. W., Koseki, H., Brockdorff, N., Ponting, C. P., Kessler, B. M., & Klose, R. J. (2012). KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. *ELife*, *2012*(1), 1–26. https://doi.org/10.7554/eLife.00205

Ferrai, C., Torlai Triglia, E., Risner-Janiczek, J. R., Rito, T., Rackham, O. J., de Santiago, I., Kukalev, A., Nicodemi, M., Akalin, A., Li, M., Ungless, M. A., & Pombo, A. (2017). RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. *Molecular Systems Biology*, *13*(10), 946. https://doi.org/10.15252/msb.20177754

Fitz, J., Neumann, T., Steininger, M., Wiedemann, E., Garcia, A. C., Athanasiadis, A., Schoeberl, U. E., & Pavri, R. (2020). Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction Johanna. *Nature Genetics*, *52*, 505–515. https://doi.org/10.1038/s41588-020-0605-6

REFERENCES

Frangini, A., Sjöberg, M., Roman-Trufero, M., Dharmalingam, G., Haberle, V., Bartke, T., Lenhard, B., Malumbres, M., Vidal, M., & Dillon, N. (2013). The Aurora B Kinase and the Polycomb Protein Ring1B Combine to Regulate Active Promoters in Quiescent Lymphocytes. *Molecular Cell*, *51*(5), 647–661. https://doi.org/10.1016/j.molcel.2013.08.022

Franke, M., Calle-mustienes, E. De, Neto, A., Acemel, R. D., & Juan, J. (2020). *CTCF knockout in zebrafish induces alterations in regulatory landscapes and developmental gene expression*. 1–39.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, *466*(7305), 490–493. https://doi.org/10.1038/nature09158

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085

Galupa, R., & Crocker, J. (2020). Enhancer–Promoter Communication: Thinking Outside the TAD. *Trends in Genetics*, *36*(7), 459–461. https://doi.org/10.1016/j.tig.2020.04.002

Gao, Z., Lee, P., Stafford, J. M., von Schimmelmann, M., Schaefer, A., & Reinberg, D. (2014). An AUTS2–Polycomb complex activates gene expression in the CNS. *Nature*, *516*(7531), 349–354. https://doi.org/10.1038/nature13921

Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate Genomes. *Journal of Molecular Biology*, *196*, 261–282. https://doi.org/10.1016/S0040-4020(02)00396-4

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., & Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics*, *51*(8), 1272–1282. https://doi.org/10.1038/s41588-019-0462-3

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Grand, R. S., Burger, L., Gräwe, C., Michael, A. K., Isbel, L., Hess, D., Hoerner, L., Iesmantavicius, V., Durdu, S., Pregnolato, M., Krebs, A. R., Smallwood, S. A., Thomä, N., & Vermeulen, M. (2021). *BANP opens chromatin and activates CpG-island-regulated genes*. *August 2020*. https://doi.org/10.1038/s41586-021-03689-8

Gröschel, S., Sanders, M. A., Hoogenboezem, R., De Wit, E., Bouwman, B. A. M., Erpelinck, C., Van Der Velden, V. H. J., Havermans, M., Avellino, R., Van Lom, K., Rombouts, E. J., Van Duin, M., Döhner, K., Beverloo, H. B., Bradner, J. E., Döhner, H., Löwenberg, B., Valk, P. J. M., Bindels, E. M. J., … Delwel, R. (2014). A single oncogenic enhancer rearrangement causes concomitant EVI1

and GATA2 deregulation in Leukemia. *Cell*, *157*(2), 369–381. https://doi.org/10.1016/j.cell.2014.02.019

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, *130*(1), 77–88. https://doi.org/10.1016/j.cell.2007.05.042

Haberle, V., Arnold, C. D., Pagani, M., Rath, M., Schernhuber, K., & Stark, A. (2019). Transcriptional cofactors display specificity for distinct types of core promoters. *Nature*, *570*(7759), 122–126. https://doi.org/10.1038/s41586-019-1210-7

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 1–17. https://doi.org/10.1038/s41580-018-0028-8

Habibi, E., Brinkman, A. B., Arand, J., Kroeze, L. I., Kerstens, H. H. D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N. C., Benedetti, R., Altucci, L., Jansen, J. H., Walter, J., Gut, I. G., Marks, H., & Stunnenberg, H. G. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, *13*(3), 360–369. https://doi.org/10.1016/j.stem.2013.06.002

Halfon, M. S. (2019). Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends in Genetics*, *35*(2), 93–103. https://doi.org/10.1016/j.tig.2018.11.004

Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Biggs, D., Preece, C., Downes, D. J., Gosden, M., Sharpe, J. A., Sloane-Stanley, J. A., Hughes, J. R., Davies, B., & Higgs, D. R. (2017). Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature Cell Biology*, *19*(8), 952–961. https://doi.org/10.1038/ncb3573

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., & Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. *Nature Cell Biology*, *15*(4), 373–384. https://doi.org/10.1038/ncb2702

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., & Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. https://doi.org/10.1038/ng1966

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

REFERENCES

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes and Development*, *32*(1), 26–41. https://doi.org/10.1101/gad.309351.117

Ho, Y., Elefant, F., Liebhaber, S. A., & Cooke, N. E. (2006). Locus Control Region Transcription Plays an Active Role in Long-Range Gene Activation. *Molecular Cell*, *23*(3), 365–375. https://doi.org/10.1016/j.molcel.2006.05.041

Hsieh, C. L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., Sun, T., Sweeney, C. J., Lee, G. S. M., Chen, S., Balk, S. P., Liu, X. S., Brown, M., & Kantoff, P. W. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), 7319–7324. https://doi.org/10.1073/pnas.1324151111

Hu, D., Gao, X., Cao, K., Morgan, M. A., Mas, G., Smith, E. R., Volk, A. G., Bartom, E. T., Crispino, J. D., Di Croce, L., & Shilatifard, A. (2017). Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. *Molecular Cell*, *65*(3), 460-475.e6. https://doi.org/10.1016/j.molcel.2017.01.013

Ibrahim, D. M., & Mundlos, S. (2020). The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. *Current Opinion in Genetics and Development*, *61*, 1–8. https://doi.org/10.1016/j.gde.2020.02.015

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., & Bird, A. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biology*, *6*(1), 0037–0051. https://doi.org/10.1371/journal.pbio.0060022

Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R., & Bird, A. P. (2010). Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, *6*(9), e1001134. https://doi.org/10.1371/journal.pgen.1001134

Isono, K., Endo, T., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B., & Koseki, H. (2013). SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. *Developmental Cell*, *26*(6), 565–577. https://doi.org/10.1016/j.devcel.2013.08.016

Jaeger, M. G., Schwalb, B., Mackowiak, S. D., Velychko, T., Hanzl, A., Imrichova, H., Brand, M., Agerer, B., Chorn, S., Nabet, B., Ferguson, F. M., Müller, A. C., Bergthaler, A., Gray, N. S., Bradner, J. E., Bock, C., Hnisz, D., Cramer, P., & Winter, G. E. (2020). Selective Mediator dependence of cell-type-specifying transcription. *Nature Genetics*, *52*, 719–727. https://doi.org/10.1038/s41588-020-0635-0

REFERENCES

Jaeger, S. A., Chan, E. T., Berger, M. F., Stottmann, R., Hughes, T. R., & Bulyk, M. L. (2010). Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. *Genomics*, *95*(4), 185–195. https://doi.org/10.1016/j.ygeno.2010.01.002

Kim, Y. J., Björklund, S., Li, Y., Sayre, M. H., & Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*, *77*(4), 599–608. https://doi.org/10.1016/0092-8674(94)90221-6

Kondo, T., Isono, K., Kondo, K., Endo, T. A., Itohara, S., Vidal, M., & Koseki, H. (2014). Polycomb Potentiates Meis2 Activation in Midbrain by Mediating Interaction of the Promoter with a Tissue-Specific Enhancer. *Developmental Cell*, *28*(1), 94–101. https://doi.org/10.1016/j.devcel.2013.11.021

Kragesteen, B. K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A. M., Jerković, I., Harabula, I., Guckelberger, P., Pechstein, M., Wittler, L., Chan, W., Franke, M., Lupiáñez, D. G., Kraft, K., Timmermann, B., … Andrey, G. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis ,. *Nature Genetics*, *50*(October), 1463–1473. https://doi.org/10.1038/s41588-018-0221-x

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Kvon, E. Z., Waymack, R., Elabd, M. G., & Wunderlich, Z. (2008). Enhancer redundancy in development and disease. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-00311-x

Kvon, E. Z., Waymack, R., Elabd, M. G., & Wunderlich, Z. (2021). Enhancer redundancy in development and disease. *Nature Reviews Genetics*, *22*(May). https://doi.org/10.1038/s41576-020-00311-x

Lagha, M., Bothma, J. P., & Levine, M. (2012). Mechanisms of transcriptional precision in animal development. *Trends in Genetics*, *28*(8), 409–416. https://doi.org/10.1016/j.tig.2012.03.006

Landolin, J. M., Johnson, D. S., Trinklein, N. D., Aldred, S. F., Medina, C., Shulha, H., Weng, Z., & Myers, R. M. (2010). Sequence features that drive human promoter function and tissue specificity. *Genome Research*, *20*(7), 890–898. https://doi.org/10.1101/gr.100370.109

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Laugsch, M., Bartusel, M., Rehimi, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., Zentis, P., Nikolic, M., Bleckwehl, T., Kolovos, P., van Ijcken, W. F. J., Šarić, T., Koehler, K., Frommolt, P., Lachlan, K., Baptista, J., & Rada-Iglesias, A. (2019). Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs. *Cell Stem Cell*, *24*(5), 736-752.e12. https://doi.org/10.1016/j.stem.2019.03.004

REFERENCES

Lee, K. K., & Workman, J. L. (2007). Histone acetyltransferase complexes: One size doesn't fit all. *Nature Reviews Molecular Cell Biology*, *8*(4), 284–295. https://doi.org/10.1038/nrm2145

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, *13*(4), 233–245. https://doi.org/10.1038/nrg3163

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, *20*(17), R754–R763. https://doi.org/10.1016/j.cub.2010.06.070

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H. S., Glass, C. K., & Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, *498*(7455), 516–520. https://doi.org/10.1038/nature12210

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., & Liu, X. S. (2011). Cistrome: An integrative platform for transcriptional regulation studies. *Genome Biology*, *12*(8), R83. https://doi.org/10.1186/gb-2011-12-8-r83

Lloret-Llinares, M., Karadoulama, E., Chen, Y., Wojenski, L. A., Villafano, G. J., Bornholdt, J., Andersson, R., Core, L., Sandelin, A., & Jensen, T. H. (2018). The RNA exosome contributes to gene expression regulation during stem cell differentiation. *Nucleic Acids Research*, *46*(21), 11502–11513. https://doi.org/10.1093/nar/gky817

Long, H. K., Blackledge, N. P., & Klose, R. J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society Transactions*, *41*, 727–740. https://doi.org/10.1042/BST20130028

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, *167*(5), 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018

Long, H. K., Sims, D., Heger, A., Blackledge, N. P., Kutter, C., Wright, M. L., Gr??tzner, F., Odom, D. T., Patient, R., Ponting, C. P., & Klose, R. J. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *ELife*, *2*, 1–19. https://doi.org/10.7554/eLife.00348

Loubiere, V., Papadopoulos, G. L., Szabo, Q., Martinez, A. M., & Cavalli, G. (2020). Widespread activation of developmental gene expression characterized by PRC1-dependent chromatin looping. *Science Advances*, *6*(2), eaax4001. https://doi.org/10.1126/sciadv.aax4001

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M.,

REFERENCES

Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Lupiáñez, D. G., Spielmann, M., & Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*, *32*(4), 225–237. https://doi.org/10.1016/j.tig.2016.01.003

Mateo, L. J., Murphy, S. E., Hafner, A., Cinquini, I. S., Walker, C. A., & Boettiger, A. N. (2019). Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, *568*(7750), 49–54. https://doi.org/10.1038/s41586-019-1035-4

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., Dsouza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., … Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257. https://doi.org/10.1038/nature09165

McKeon, C., Ohkubo, H., Pastan, I., & Crombrugghe, B. de. (1982). Unusual methylation pattern of the α2(I) collagen gene. *Cell*, *29*(1), 203–210. https://doi.org/10.1016/0092-8674(82)90104-0

Mehrjouy, M. M., Fonseca, A. C. S., Ehmke, N., Paskulin, G., Novelli, A., Benedicenti, F., Mencarelli, M. A., Renieri, A., Busa, T., Missirian, C., Hansen, C., Abe, K. T., Speck-Martins, C. E., Vianna-Morgante, A. M., Bak, M., & Tommerup, N. (2018). Regulatory variants of FOXG1 in the context of its topological domain organisation /631/208/200 /631/208/1516 article. *European Journal of Human Genetics*, *26*(2), 186–196. https://doi.org/10.1038/s41431-017-0011-4

Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Vrielink, J. A. F. O., Elkon, R., Melo, S. A., Léveillé, N., Kalluri, R., de Laat, W., & Agami, R. (2013). ERNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Molecular Cell*, *49*(3), 524–535. https://doi.org/10.1016/j.molcel.2012.11.021

Merkenschlager, M., & Nora, E. P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu. Rev. Genom. Hum. Genet.*, *17*, 17–43. https://doi.org/10.1146/annurev-genom-083115-022339

Mir, M., Bickmore, W., Furlong, E. E. M., & Narlikar, G. (2019). *Chromatin topology , condensates and gene regulation : shifting paradigms or just a phase ?* 1–6. https://doi.org/10.1242/dev.182766

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., De Laat, W., Spitz, F., & Duboule, D. (2011). A regulatory archipelago controls hox genes transcription in digits. *Cell*, *147*(5), 1132–1145. https://doi.org/10.1016/j.cell.2011.10.023

REFERENCES

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G. L., & Sartorelli, V. (2013). ERNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. *Molecular Cell*, *51*(5), 606–617. https://doi.org/10.1016/j.molcel.2013.07.022

Muller, M., & Lindmark, D. G. (1976). Uptake of Metronidazole and Its Effect on Viability in Trichomonads and Entamoeba invadens Under Anaerobic and Aerobic Conditions. *Antimicrobial Agents and Chemotherapy*, *9*(4), 696–700. https://doi.org/10.1128/AAC.9.4.696

Neumayr, C., Pagani, M., Stark, A., & Arnold, C. D. (2019). STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Current Protocols in Molecular Biology*, *128*(1), e105. https://doi.org/10.1002/cpmb.105

Nora, E. P., Goloborodko, A., Valton, A., Dekker, J., Mirny, L. A., Bruneau, B. G., Nora, P., Goloborodko, A., Valton, A., Gibcus, J. H., Uebersohn, A., & Abdennur, N. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930–944. https://doi.org/10.1016/j.cell.2017.05.004

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*, 381–385. https://doi.org/10.1038/nature11049

Osterwalder, M., Barozzi, I., Tissiéres, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691), 239–243. https://doi.org/10.1038/nature25461

Pachano, T., Crispatzu, G., & Rada-iglesias, A. (2019). Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Briefings in Functional Genomics*, *18*(6), 358–366. https://doi.org/10.1093/bfgp/elz022

Pachano, T., Crispatzu, G., & Rada-Iglesias, A. (2019). Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Briefings in Functional Genomics*, *18*(6). https://doi.org/10.1093/bfgp/elz022

Pagès, H. (2020). *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.56.0.*

Paralkar, V. R., Taborda, C. C., Huang, P., Yao, Y., Kossenkov, A. V., Prasad, R., Luan, J., Davies, J. O. J., Hughes, J. R., Hardison, R. C., Blobel, G. A., & Weiss, M. J. (2016). Unlinking an lncRNA from

Its Associated cis Element. *Molecular Cell*, *62*(1), 104–110. https://doi.org/10.1016/j.molcel.2016.02.029

Perino, M., Mierlo, G. Van, Karemaker, I. D., Genesen, S. Van, Vermeulen, M., Marks, H., Heeringen, S. J. Van, & Veenstra, G. J. C. (2018). MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. *Nature Genetics*, *50*(July), 1002–1010. https://doi.org/10.1038/s41588-018-0134-8

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of drosophila gastrulation. *Current Biology*, *20*(17), 1562–1567. https://doi.org/10.1016/j.cub.2010.07.043

Perry, M. W., Boettiger, A. N., & Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13570–13575. https://doi.org/10.1073/pnas.1109873108

Pnueli, L., Rudnizky, S., Yosefzon, Y., & Melamed, P. (2015). RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin α-subunit gene. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(14), 4369–4374. https://doi.org/10.1073/pnas.1414841112

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., … Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, *515*(7527), 402–405. https://doi.org/10.1038/nature13986

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. a, Flynn, R. a, & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283. https://doi.org/10.1038/nature09692

Ramirez-Carrozzi, V. R., Braas, D., Bhatt, D. M., Cheng, C. S., Hong, C., Doty, K. R., Black, J. C., Hoffmann, A., Carey, M., & Smale, S. T. (2009). A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. *Cell*, *138*(1), 114–128. https://doi.org/10.1016/j.cell.2009.04.020

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*, W160–W165. https://doi.org/10.1093/nar/gkw257

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., … Aiden, E. L.

(2017a). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., … Aiden, E. L. (2017b). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., Hanscom, C., Pillalamarri, V., Seabra, C. M., Abbott, M. A., Abdul-Rahman, O. A., Aberg, E., Adley, R., Alcaraz-Estrada, S. L., Alkuraya, F. S., An, Y., Anderson, M. A., Antolik, C., Anyane-Yeboa, K., … Talkowski, M. E. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, *49*(1), 36–45. https://doi.org/10.1038/ng.3720

Rehimi, R., Nikolic, M., Cruz-Molina, S., Tebartz, C., Frommolt, P., Mahabir, E., Clément-Ziza, M., & Rada-Iglesias, A. (2016). Epigenomics-Based Identification of Major Cell Identity Regulators within Heterogeneous Cell Populations. *Cell Reports*, *17*(11), 3062–3076. https://doi.org/10.1016/j.celrep.2016.11.046

Reiter, F., Wienerroither, S., & Stark, A. (2017). Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics and Development*, *43*, 73–81. https://doi.org/10.1016/j.gde.2016.12.007

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, *19*(12), 789–800. https://doi.org/10.1038/s41576-018-0060-8

Sabari, B. R., Dall, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V, Manteiga, J. C., Li, C. H., Guo, Y. E., Day, D. S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Ihn Lee, T., Cisse, I. I., … Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, *361*, eaar3958. https://doi.org/10.1126/science.aar3958

Saksouk, N., Avvakumov, N., Champagne, K. S., Hung, T., Doyon, Y., Cayrou, C., Paquet, E., Ullah, M., Landry, A. J., Côté, V., Yang, X. J., Gozani, O., Kutateladze, T. G., & Côté, J. (2009). HBO1 HAT Complexes Target Chromatin throughout Gene Coding Regions via Multiple PHD Finger Interactions with Histone H3 Tail. *Molecular Cell*, *33*(2), 257–265. https://doi.org/10.1016/j.molcel.2009.01.007

Sanborn, A. L., Rao, S. S. P., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., & Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National*

Academy of Sciences of the United States of America, *112*(47), E6456–E6465. https://doi.org/10.1073/pnas.1518552112

Schaffner, W. (2015). Enhancers, enhancers - From their discovery to today's universe of transcription enhancers. *Biological Chemistry*, *396*(4), 311–327. https://doi.org/10.1515/hsz-2014-0303

Schmitt, A. D., Hu, M., & Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Publishing Group*, *17*(12), 743–755. https://doi.org/10.1038/nrm.2016.104

Schröder, S., Herker, E., Itzen, F., He, D., Thomas, S., Gilchrist, D. A., Kaehlcke, K., Cho, S., Pollard, K. S., Capra, J. A., Schnölzer, M., Cole, P. A., Geyer, M., Bruneau, B. G., Adelman, K., & Ott, M. (2013). Acetylation of RNA Polymerase II Regulates Growth-Factor-Induced Gene Transcription in Mammalian Cells. *Molecular Cell*, *52*(3), 314–324. https://doi.org/10.1016/j.molcel.2013.10.009

Serebreni, L., & Stark, A. (2021). Insights into gene regulation: From regulatory genomic elements to DNA-protein and protein-protein interactions. *Current Opinion in Cell Biology*, *70*, 58–66. https://doi.org/10.1016/j.ceb.2020.11.009

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, *15*(4), 272–286. https://doi.org/10.1038/nrg3682

Smith, E., & Shilatifard, A. (2014). Enhancer biology and enhanceropathies. *Nature Structural and Molecular Biology*, *21*(3), 210–219. https://doi.org/10.1038/nsmb.2784

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews. Genetics*, *19*(7), 453–467. https://doi.org/10.1038/s41576-018-0007-0

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.*, *13*(9), 613–626. https://doi.org/10.1038/nrg3207

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.-J., Wendt, K. S., Grosveld, F., van Ijcken, W., & Soler, E. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature Protocols*, *8*(3), 509–524. https://doi.org/10.1038/nprot.2013.018

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., & Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, *480*(7378), 490–495. https://doi.org/10.1038/nature10716

Stanojevic, D., Small, S., & Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science*, *254*(5036), 1385–1387. https://doi.org/10.1126/science.1683715

REFERENCES

Stein, R., Sciaky-Gallili, N., Razin, A., & Cedar, H. (1983). Pattern of methylation of two genes coding for housekeeping functions. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(9 I), 2422–2426. https://doi.org/10.1073/pnas.80.9.2422

Tsukada, Y. I., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P., & Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, *439*(7078), 811–816. https://doi.org/10.1038/nature04433

van Ingen, H., van Schaik, F. M. A., Wienk, H., Ballering, J., Rehmann, H., Dechesne, A. C., Kruijzer, J. A. W., Liskamp, R. M. J., Timmers, H. T. M., & Boelens, R. (2008). Structural Insight into the Recognition of the H3K4me3 Mark by the TFIID Subunit TAF3. *Structure*, *16*(8), 1245–1256. https://doi.org/10.1016/j.str.2008.04.015

Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W. M. P., van Schaik, F. M. A., Varier, R. A., Baltissen, M. P. A., Stunnenberg, H. G., Mann, M., & Timmers, H. T. M. (2007). Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, *131*(1), 58–69. https://doi.org/10.1016/j.cell.2007.08.016

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., & Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. https://doi.org/10.1038/nature07730

Wu, X., Johansen, J. V., & Helin, K. (2013). Fbxl10/Kdm2b Recruits Polycomb Repressive Complex 1 to CpG Islands and Regulates H2A Ubiquitylation. *Molecular Cell*, *49*(6), 1134–1146. https://doi.org/10.1016/j.molcel.2013.01.016

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., & Allis, C. D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, *442*(7098), 86–90. https://doi.org/10.1038/nature04815

Xiao, J., Hafner, A., & Boettiger, A. N. (2020). How subtle changes in 3D structure can create large changes in transcription. *BioRxiv*, 1–25. https://doi.org/10.1101/2020.10.22.351395

Yao, X., Zhang, M., Wang, X., Ying, W., Hu, X., Dai, P., Meng, F., Shi, L., Sun, Y., Yao, N., Zhong, W., Li, Y., Wu, K., Li, W., Chen, Z., & Yang, H. (2018). Tild-CRISPR Allows for Efficient and Precise Gene Knockin in Mouse and Human Cells. *Developmental Cell*, *45*(4), 526-536.e5. https://doi.org/10.1016/J.DEVCEL.2018.04.021

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556–559. https://doi.org/10.1038/nature13994

REFERENCES

Zamudio, A. V, Agnese, A. D., Henninger, J. E., Taatjes, D. J., Schuijers, J., Young, R. A., Zamudio, A. V, Agnese, A. D., Henninger, J. E., Manteiga, J. C., Afeyan, L. K., Hannett, N. M., Coffey, E. L., Li, C. H., Oksuz, O., Sabari, B. R., Boija, A., & Klein, I. A. (2019). Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes Article Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. *Molecular Cell*, *76*(5), 753-766.e6. https://doi.org/10.1016/j.molcel.2019.08.016

Zentner, G. E., Tesar, P. J., Scacheri, P. C., Zentner, G. E., Tesar, P. J., & Scacheri, P. C. (2011). *Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions*. 1273–1283. https://doi.org/10.1101/gr.122382.111

Zhen, C. Y., Tatavosian, R., Huynh, T. N., Duc, H. N., Das, R., Kokotovic, M., Grimm, J. B., Lavis, L. D., Lee, J., Mejia, F. J., Li, Y., Yao, T., & Ren, X. (2016). Live-cell single-molecule tracking reveals co-recognition of H3K27me3 and DNA targets polycomb Cbx7-PRC1 to chromatin. *ELife*, *5*(OCTOBER2016), 1–36. https://doi.org/10.7554/eLife.17667.001

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Meister, P., Smallwood, S., & Giorgetti, L. (2021). Nonlinear control of transcription through enhancer-promoter interactions. *BioRxiv*, 2021.04.22.440891.

Zylicz, J. J., Dietmann, S., Günesdogan, U., Hackett, J. A., Cougot, D., Lee, C., & Surani, M. A. (2015). Chromatin dynamics and the role of G9a in gene regulation and enhancer silencing during early mouse development. *ELife*, *4*(NOVEMBER2015), 1–25. https://doi.org/10.7554/eLife.09571

FIGURE INDEX

FIGURE INDEX

7- TABLE INDEX

LIST OF ABBREVIATIONS

3C	Chromatin conformation capture
4C-seq	Chromatin conformation capture combined with high-throughput sequencing
AntNPC	Anterior neural progenitor cells
ATAC-seq	Assay for transposable accessible chromatin-sequencing
ATP	Adenosine triphosphate
BOFS	Branchiooculofacial syndrom
bp	Base pair
BSA	Bovine serum ALbumina
C2H3NaO2	Sodium Acetate
CAP	CxxC affinity purification
cDNA	Complementary DNA
CGI	CpG island
ChIP-qPCR	Chromatin immunoprecipitation followed by quantitative polymerase chain reaction
ChIP-seq	Chromatin immunoprecipitation sequencing
CRISPR	Clustered regulatory interspaced short palindromic repeats
CTCF	CCCTC-binding factor
CTD	Carboxy-terminal domain
CxxC	Zinc finger CxxC domain-containing
DAPI	4,6-diamidin-2-phenylindol
DMEM	Dulbeccos Modified Eagles Medium
DNA	Desoxiribonucleic acid
DNTP	Desoxyribonucleotide triphosphate
DPE	Downstream promoter element

LIST OF ABBREVIATIONS

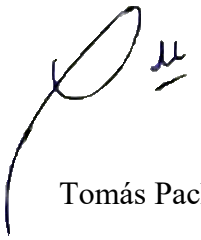
DTT	Dithiothreitol
EDTA	ethylendiaminetetraacetic acid
EGTA	thylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid)
eRNA	Enhancer ribonucleic acid
ESC	Embryonic stem cell
FBS	Fetal bovine serum
FPKM	Fragmentes per kilobase of exon model per million reads mapped
gRNA	guideRNA
H2AK119ub	Histone 2A lysine 119 ubiquitination
H3K27ac	Histone 3 lysine 27 acetylation
H3K36me2	Histone 3 lysine 36 double methylation
H3K4me1	Histone 3 lysine 4 mono methylation
H3K4me3	Histone 3 lysine 4 triple methylation
HCl	Hydrochloric acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hESC	Human embryonic stem cell
IF	Immunofluorece
KCl	Potasium Chloride
LB1	Lysis Buffer 1
LB2	Lysis Buffer 2
LB3	Lysis Buffer 3
LiCl	Litium Chloride
LIF	Leukemia inhibitory factor
mESC	Mouse embryonic stem cell
MgCl2	Magnesium Chloride
mRNA	Messenger ribonucleic acid
NaCl	Sodium Chloride

LIST OF ABBREVIATIONS

NCC	Neural crest cells
NEB	New England Biolab
oCGI	orphan CpG island
PcG	Polycomb group complex
pCGI	Promoter-associated CpG island
PCR	Polymerase chain reaction
PE	Poised enhancer
PIC	Pre-initiation complex-dependent
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
RIPA	Radio immunoprecipitation assay buffer
RNA	Ribonucleic acid
RNAP2	RNA polymerase II
RT-qPCR	Real-time quantitative polymerase chain reaction
SDS	Sodium dodecyl sulfate
STAP-seq	Self-transcribing active promoter sequencing
STARR-seq	Self-transcribing active regulatory region sequencing
TADs	Topological associated domains
TFBS	Transcription factor binding site
TFs	Transcription factors
TSS	Transcription start site

EIDESSTATTLICHE ERKLÄRUNG

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstandig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit −einschliesslich Tabellen, Karten und Abbildungen −, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie − abgesehen von unten angegebenen Teilpublikationen − noch nicht veroÅNffentlicht worden ist sowie, dass ich eine solche Veroöffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Dr. Alvaro Rada-Iglesias, Prof. Dr. Mirka Uhlirova, Dr. Peter Tessarz und Prof. Dr. Siegfried Roth betreut worden.



Köln in October 2021 Tomás Pachano

Nachfolgende Teilpublikationen liegen vor:

Pachano, T., Sánchez-Gaya, V., Ealo, T. *et al.* Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nature Genetics* 53, 1036–1049 (2021). https://doi.org/10.1038/s41588-021-00888-x.

CURRICULUM VITAE

Tomás Pachano

Birth	25th of January 1992 Buenos Aires, Argentina
Address	Camilo Alonso Vega 30B, 39007, Santander, Cantabria, Spain
Contact	t.pachano@uni-koeln.de / tomas.pachano@unican.es / pachano.tomas@gmail.com

Academic history

2017 –	Cologne University, Cologne, Germany Dr. Rer. Nat., Graduate School for Biological Sciences Doctoral Thesis: "Dissecting the role of orphan CpG islands in dictating compatibility between of poised enhancers and developmental genes"
2011 – 2016	National University of San Martin, Buenos Aires, Argentina Diploma Studies in Agrobiotechnological Engineer. Degree average: 8.82/10 Diploma thesis: "Role of histone acetylation in transcription regulation and pathogenesis in *Trichomonas vaginalis*"

Academic scholarships

2017 – 2019	DAAD Research Grants - Doctoral Program in Germany 2017-2021 for the total of my doctoral studies
2011 – 2016	National University of San Martin scholarship grant for the total of diploma studies

CURRICULUM VITAE

Research

2021 - 2021	**Developmental genomics Lab. Head director: Dr. Alvaro Rada-Iglesias. Instituto de Biomedicina y Biologia Molecular de Cantabria (IBBTEC, University of Cantabria)**
2017 - 2021	**Developmental genomics Lab. Head director: Dr. Alvaro Rada-Iglesias. Center for Molecular Medicine Cologne (CMMC, University of Cologne)**
2014 - 2016	**Anaerobic Parasites Lab. Head director: Dr. Natalia de Miguel. Instituto de Investigaciones Biotecnológicas – Instituto Tecnológico de Chascomús (IIB-INTECH, CONICET-UNSAM)**
2013 - 2014	**Developmental Biology Lab. Head director: Dr. Pablo Strobl-Mazzulla. Instituto de Investigaciones Biotecnológicas – Instituto Tecnológico de Chascomús (IIB-INTECH, CONICET-UNSAM)**

Scientific Publications

Peer-reviewed publications

1. **2021.** "ORPHAN CPG ISLANDS AMPLIFY POISED ENHANCER REGULATORY ACTIVITY AND DETERMINE TARGET GENE RESPONSIVENESS". **Pachano, T**; Sánchez-Gaya, V; Ealo, T; Mariner-Faulí, M; G. Asenjo, H, Respuela, P; Cruz-Molina, S; Muñoz-San Martín, M; Haro, E; van Ijcken, W; Landeira, D; Rada-Iglesias, A. ***Nature Genetics***.
2. **2021.** "THE CHROMATIN, TOPOLOGICAL AND REGULATORY PROPERTIES OF PLURIPOTENCY-ASSOCIATED POISED ENHANCERS ARE CONSERVED IN VIVO". Crispatzu, G; Rehimi, R; **Pachano, T**; Bleckwehl, T; Cruz-Molina, S; Xiao, C; Mahabir, E; Bazzi, H; Rada-Iglesias, A. ***Nature Communications***.
3. **2019.** "POLYCOMB PROTEINS AS ORGANIZERS OF 3D GENOME ARCHITECTURE IN EMBRYONIC STEM CELLS". **Pachano, T**; Crispatzu, G; Rada-Iglesias, A. ***Brieflings in Functional Genomics***.
4. **2016.** "EPIGENETIC REGULATES TRANSCRIPTION AND PATHOGENESIS IN THE PARASITE TRICHOMONAS VAGINALIS". **Pachano, T**; Nievas, Y.R; Lizarraga, A; Johnson, P.J; Strobl-Mazzulla, P.H.; de Miguel, N. ***Cellular Microbiology***.

Conference presentations

1. **2020**. “ORPHAN CPG ISLANDS DICTATE THE COMPATIBILITY BETWEEN POISED ENHANCERS AND DEVELOPMENTAL GENES”. **Pachano, T**; Sánchez-Gaya, V; Mariner-Faulí, M; Ealo, T; Respuela, P; G. Asenjo; Landeira, D; Rada-Iglesias, A. XVII Spanish Society for Developmental Biology Meeting. Bilbao, Spain (Virtual). **Talk**
2. **2020**. “DISSECTING THE ROLE OF CPG ISLANDS AS REGULATORY BOOSTERS OF POISED ENHANCERS”. **Pachano, T**; Sánchez-Gaya, V; Mariner-Faulí, M; Ealo, T; G. Asenjo; H, Respuela, P; Landeira, D; Rada-Iglesias, A. V Cold Spring Harbor Laboratoy meeting on Epigenetics and Chromatin. Cold Spring Harbor, USA (Virtual). **Poster**
3. **2019**. “DISSECTING THE ROLE OF CPG ISLANDS AS REGULATORY BOOSTERS OF POISED ENHANCERS”. **Pachano T**.; Rada-Iglesias, A. X Conference of the Latin American Society for Developmental Biology. Buenos Aires, Argentine. **Poster**
4. **2015. “**NUEVOS ASPECTOS INVOLUCRADOS EN LA PATOGÉNESIS DE TRICHOMONAS VAGINALIS”. de Miguel N.; **Pachano T**., Lizarraga A.; Nievas Y.R.; Coceres V., Strobl-Mazzulla P. VII Congreso Argentino de Parasitología. San Carlos de Bariloche, Argentine. **Plenary Session**

5. **2014. "**EPIGENETIC REGULATION OF GENES IMPLICATED IN ADHESION TO HOST CELLS IN *TRICHOMONAS VAGINALIS”.* **Pachano T**.; Villarruel F.; Lizarraga A.; Coceres V.M.; Strobl-Mazzulla P.H.; de Miguel N. X Conference of Protozoology and Parasitic Diseases. Mar del Plata, Argentine. **Poster**