Title
Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization

Author
Denholtz, Matthew

Publication Date
2013

UNIVERSITY OF CALIFORNIA

Los Angeles

Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Molecular Biology

by

Matthew Brian Denholtz

2013

ABSTRACT OF THE DISSERTATION

Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization

By

Matthew Denholtz
Doctor of Philosophy in Molecular Biology
University of California, Los Angeles, 2013
Professor Kathrin Plath, Chair

The spatial organization of the genome is linked to its biological function. However, the relationship between specific gene regulatory networks that govern cell identity and large-scale organization of genomes remains unclear. To investigate the basis of distal chromatin interactions occurring between genomic regions mega-bases away on the same or different chromosomes, we mapped long-range chromatin interactions in embryonic stem cells (ESCs) using 4C-seq, and examined the genomic features of the interacting regions. Confirming previous results, we show that open, accessible versus closed chromatin character is the primary determinant of distal chromatin interaction preferences, where interacting regions exhibit very similar open/closed chromatin character. We extend these results by demonstrating that genomic regions highly enriched for binding by the pluripotency transcription factors Oct4, Sox2, and Nanog preferentially co-localize, as do regions strongly enriched for Polycomb proteins and

trimethylation of histone H3 at lysine 27 (H3K27me3), which include the *Hox* clusters. Consistent with a spatial segregation of these transcriptional networks, we find that Nanog and Polycomb proteins occupy distinct spaces in the nucleus. Importantly, loss of the Polycomb protein Eed and H3K27me3 diminishes the preferential interactions between regions normally highly enriched for Polycomb proteins and H3K27me3 without dramatically changing long-range chromatin interactions related to the open/closed chromatin state. Finally, a comparison of interactomes in ESCs and fibroblasts uncovered an ESC-specific spatial organization to the mouse genome that is gradually re-established upon reprogramming to induced pluripotent stem cells (iPSCs). Together, our data suggest that transcriptional networks that govern ESC identity play a role in determining genome-organization. We propose the existence of a hierarchy in the organization of chromatin contacts wherein, at the largest scale, the open/closed chromatin character defines an interaction space and overall chromosome conformation; on a finer scale, cell type-specific transcriptional networks direct preferential distal interactions, which we speculate are critical for efficient regulation of transcription and establishment of local chromatin environments.

The dissertation of Matthew Brian Denholtz is approved.

Michael Carey

Siavash Kurdistani

Matteo Pellegrini

Stephen Smale

Kathrin Plath, Committee Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

FIGURES

TABLES

ACKNOWLEDGEMENTS

Funding

My tuition and stipend were supported by pre-doctoral fellowships from the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, the California Institute for Regenerative Medicine (CIRM), the UCLA Graduate Division, and by grants to KP who is supported by the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, NIH (DP2OD001686 and P01GM099134), and CIRM (RN1-00564).

Published manuscript contributors (a reproduction of which is Chapter 1)

Pluripotency in 3D: Genome organization in pluripotent cells

Matthew Denholtz[1] and Kathrin Plath[1,2]

[1]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Molecular Biology Institute, Department of Biological Chemistry,

[2]Bioinformatics Interdepartmental Program, Jonsson Comprehensive Cancer Center,

at the David Geffen School of Medicine of the University of California, Los Angeles, CA

Submitted manuscript contributors (a version of which is presented here as Chapter 2)

Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization

Matthew Denholtz[1,2,4,7]*, Giancarlo Bonora[1,2,3,4,5]*, Constantinos Chronis[1,4], Erik Splinter[8], Wouter de Laat[8], Jason Ernst[1,2,3,4,6], Matteo Pellegrini[1,2,3,5,6 #], and Kathrin Plath[1,2,3,4,5,6, #]

[1]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, [2]Molecular Biology Institute, [3]Bioinformatics Interdepartmental Program, [4]Department of Biological Chemistry, [5]Department of Molecular, Cell, and Developmental Biology, [6]Jonsson Comprehensive Cancer Center, [7]Molecular Biology Interdepartmental PhD Program, [1-7] at the David Geffen School of Medicine of the University of California, Los Angeles, CA, USA, [8]Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands

*These authors contributed equally to this work

[#]Correspondence: Matteo Pellegrini (matteop@mcdb.ucla.edu) and Kathrin Plath (kplath@mednet.ucla.edu)

People

I would first and foremost thank Kathrin for her unwavering support in all matters personal and professional throughout the course of my PhD studies. Her relentless drive and focus facilitated and guided my scientific growth at every step of my graduate education. My progress and this work would not be what they are today without her efforts, and whatever professional accomplishments I achieve in the future have had their foundations laid through her.

I would next thank Giancarlo Bonora, an equal contributor to the research presented here. His intellectual contributions to experimental designs, as well as to the analysis and interpretation of data were invaluable. Additionally, analysis software for the vast majority of the computational portion of this work, although conceptualized by all those involved, was written and executed by him. The sophistication of this analysis allowed us to identify a number of important findings in the fields of genome organization and stem cell biology, findings that would have been difficult to uncover without his efforts.

Mike Carey, Siavash Kurdistani, Matteo Pellegrini, and Steve Smale have served as a remarkably supportive thesis committee and have proven invaluable in providing scientific and career advice, as well as letters of recommendation which helped me to secure funding, and I am grateful for the time and effort they have spent doing so.

Additional thanks to all past and present members of the Plath lab, who have made the lab a fun, productive, and stimulating environment to work in. In particular, thank you to Rupa Sridharan for forcing me to think and work harder, to Mike Mason for teaching me to code, to Kostas Chronis for ChIP-seq data, to Sanjeet Patel for assistance in RNA-seq analysis, to Mark Chin for providing a fresh perspective on lab life, and to Ritchie Ho for the occasional cathartic commiseration.

Thanks to Larry Jeannotte for showing me at a young age that intellectual challenges are fun.

Thanks to Andrew Arenson for his support, friendship, and guidance during a difficult time in my life that bestowed in me a sense of confidence and self-worth that I have carried with me ever since.

Thanks to Steve James for introducing me to scientific research and for sharing his passion for it.

31 years worth of thanks to my friends and family, particularly my parents who have supported me in every way possible throughout the course of my life. My efforts would amount to much less without the stability and opportunities their love and support have provided me.

Finally, I am eternally grateful for having met and married Peggy Vorwald. Our love and partnership provide the foundation for everything I do.

VITA

EDUCATION

2001-2005 Biochemistry and Molecular Biology B.S. *cum laude* (GPA: 3.38), Gettysburg College, Gettysburg, PA

AWARDS AND HONORS

July, 2012 – March, 2013	UCLA Graduate Division Dissertation Year Fellow ($20,000 towards stipend, full tuition, $500 in research and travel funding)
October, 2012	MBI Dissertation Year Award for outstanding scholarship achievement ($2500 cash prize)
July, 2011 – June, 2012	California Institute for Regenerative Medicine pre-doctoral fellowship (full stipend and tuition, $5,000 in research and travel funding)
July, 2010 - June, 2011	Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research pre-doctoral fellowship (full stipend and tuition)

PUBLICATIONS

Denholtz M*, Bonora G*, Chronis C, Splinter E, de Laat W, Ernst J, Pelligrini M, Plath K. Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization. **Cell Stem Cell**, accepted in principle. *=Equal contribution.

Denholtz M, Plath K. Pluripotency in 3D: Genome organization in pluripotent cells. **Curr. Opinion Cell Bio, 2012** Dec;24(6):793-801. Epub 2012 Nov 27.

Kaufman DR, Goudsmit J, Holterman L, Ewald BA, **Denholtz M**, Devoy C, Giri A, Grandpre LE, Heraud JM, Franchini G, Seaman MS, Havenga MJ, Barouch DH. Differential Antigen Requirements for Protection Against Systemic and Intranasal Vaccinia Virus Challenges in Mice. **J Virol. 2008** Jul;82(14):6829-37. Epub 2008 Apr 30.

Liu J, Ewald BA, Lynch DM, **Denholtz M**, Abbink P, Lemckert AA, Carville A, Mansfield KG, Havenga MJ, Goudsmit J, Barouch DH. Magnitude and phenotype of cellular immune responses elicited by recombinant adenovirus vectors and heterologous prime-boost regimens in rhesus monkeys. **J Virol. 2008** May;82(10):4844-52. Epub 2008 Mar 12.2008

Abbink P, Lemckert AA, Ewald BA, Lynch DM, **Denholtz M**, Smits S, Holterman L, Damen I, Vogels R, Thorner AR, O'Brien KL, Carville A, Mansfield KG, Goudsmit J, Havenga MJ, Barouch DH. Comparative seroprevalence and immunogenicity of six rare serotype recombinant adenovirus vaccine vectors from subgroups B and D. **J Virol. 2007** May;81(9):4654-63. Epub 2007 Feb 28.

Thorner AR, Lemckert AA, Goudsmit J, Lynch DM, Ewald BA, **Denholtz M**, Havenga MJ, Barouch DH. Immunogenicity of heterologous recombinant adenovirus prime-boost vaccine regimens is enhanced by circumventing vector cross-reactivity**. J Virol. 2006** Dec;80(24):12009-16. Epub **2006** Oct 11.

TEACHING AND MENTORING

Winter 2010 - Summer 2013 Supervision and training of three graduate and rotation students

Spring 2009 Teaching assistant, LS3: Introduction to Molecular Biology – 2 laboratories and 2 discussion sections, UCLA

Fall 2009 Teaching assistant: LS4: Genetics – 3 Discussion sections, UCLA

PROFESSIONAL PRESENTATIONS

JCCC Gene regulation monthly meeting; UCLA, Los Angeles, CA, November 2012; oral presentation

Stem Cell Symposium: Stem Cells and Cancer: Shared Paths, Different Destinations, UCLA, Los Angeles, CA; February 2nd, 2012; poster presentation

MBI Departmental retreat, UCLA Conference Center, Lake Arrowhead, CA, October, 2011; oral presentation

CIRM Grantee Meeting, San Francisco, CA, September 15th, 2011; oral presentation

Penn State Summer Symposium: Chromatin and Epigenetic regulation of Transcription, Penn State University, June, 21st - 24th, 2011; poster presentation

Biological Chemistry Departmental Retreat, UCLA, Los Angeles, CA, May 6th, 2011; poster presentation

Gene Regulation Program Meeting, UCLA, Los Angeles, CA, April 5th, 2011; oral presentation

Biological Chemistry Departmental Meeting, UCLA, Los Angeles, CA, March 12th, 2011; oral presentation

Stem Cell Symposium: Stem Cells: Basic Biology to Translational Medicine, UCLA, Los Angeles, CA; February 18th, 2011; poster presentation

Bioinformatics Departmental Retreat, UCLA, Los Angeles, CA, February 3rd, 2011; oral presentation

MBI departmental retreat, ULCA Conference Center, Lake Arrowhead, CA, October 14th, 2010; oral presentation

Biological Chemistry Departmental Retreat, UCLA, Los Angeles, CA, April 30th, 2010; poster presentation

ADDITIONAL MEETINGS ATTENDED

Keystone Symposium: *Chromatin dynamics and higher order organization*, February 25th – March 2nd, 2009, Coeur d'Alene, ID

LEADERSHIP AND COMMUNITY INVOLVEMENT

2010-2012 President, Biological Sciences Council, UCLA

2008-2011 MBI representative to the Biological Sciences Council, Graduate Student Association, UCLA

2009-2011 MBI Departmental retreat organization committee chair, UCLA

2008-2011 MBI student representative, UCLA

2009-2010 ACCESS Steering committee member, UCLA

2008-2010 Biological Sciences Council representative to the Graduate Student Association, UCLA

2009 2009 ACCESS program recruitment leader, UCLA

OVERVIEW

The spatial organization of the genome is linked to its biological function. However, the relationship between specific gene regulatory networks that govern cell identity and large-scale organization of genomes remains unclear. We initiated this work to better understand the relationship between genome organization and gene regulation, and to explore the differences in genome organization between pluripotent and differentiated cell types.

Specifically, we addressed the following questions:

1. Is there a three-dimensional organization of chromatin specific to pluripotent cells, and if so, is it re-established upon somatic cell reprogramming to the pluripotent state?
2. Can correlations between well-characterized linear genomic features and features of genome organization help us to identify mechanistic drivers of genome organization?
3. Can we perturb the organization of the genome by genetic manipulation of the putative mechanistic effectors identified in (2)?

In order to answer these questions we used circular chromosome conformation capture coupled to high throughput sequencing (4C-seq)[1] to identify genome-wide chromatin contacts made by a variety genomic regions ("bait" loci) in embryonic stem cells (ESCs, pluripotent cells derived from the inner cell mass of the pre-implantation blastocyst), induced pluripotent stem cells (iPSCs, ES-like cells generated via the ectopic expression of pluripotency transcription factors in differentiated cells), partially reprogramed cells (pre-iPSCs), and mouse embryonic fibroblasts (MEFs). These chromatin contact maps (interactomes) allowed us to identify

differences in genome organizational features between different cell types, as well as to compare the genomic features of interacting genomic regions in order to identify genomic features shared between distal chromatin elements co-localized in three-dimensional space. These features could potentially act as drivers of genome organization.

A comparison of interactomes in ESCs and fibroblasts uncovered an ESC-specific organization to the mouse genome that is gradually re-established upon reprogramming to iPSCs. Confirming previous results in mouse[2], we showed that open, accessible versus closed chromatin character is the primary determinant of distal chromatin interaction preferences, where interacting regions exhibit very similar open/closed chromatin character, a fundamental organizational feature conserved across metazoan genomes[3,4] (Chapters 2.3.3 and 2.3.6). We extend these results by showing that genomic regions highly enriched for binding by the pluripotency transcription factors Oct4, Sox2, and Nanog (OSN) preferentially co-localize, as do regions strongly enriched for Polycomb proteins and histone H3K27me3, including the *Hox* clusters (Chapter 2.3.4).

Consistent with a spatial segregation of these transcriptional networks, we show that Nanog and Polycomb proteins occupy distinct spaces in the nucleus via immunostaining and cell imaging (Chapter 2.3.5).

We further went on to show that loss of the Polycomb protein Eed and tri-methylation of histone 3 at lysine 27 (H3K27me3) diminishes the preferential interactions between regions normally highly enriched for Polycomb proteins and H3K27me3 without dramatically changing distal chromatin interactions related to the open/closed chromatin state (Chapter 2.3.7).

Together, these data suggest that transcriptional networks that govern ESC identity play a role in determining genome-organization. We propose the existence of a hierarchy in the

organization of chromatin contacts wherein, at the largest scale, open/closed chromatin character defines an interaction space and overall chromosome conformation; on a finer scale, cell type-specific transcriptional networks direct preferential distal interactions, which we speculate are critical for efficient regulation of transcription and the establishment of local chromatin environments.

Chapter 1 is a reproduction (with permission from the publishers, and with minor changes in formatting, referencing, and figure layout to accommodate dissertation structure) of an invited review written by Kathrin Plath and I, published in 2012 in *Current Opinions in Cell Biology*, entitled "Pluripotency in 3D: Genome organization in pluripotent cells"[5]. It discusses the recent publications in the field of genome organization, with a focus on genome organization in pluripotent cells and changes in genome organization upon differentiation of pluripotent cells.

Chapter 2 is a modified version of a manuscript under review at *Cell Stem Cell* as of August 1st, 2013. This manuscript is the culmination of the work I have conducted as a graduate student in Kathrin Plath's lab. Distinct from the general nature of the introduction in Chapter 1, the introduction to Chapter 2 frames the specific questions addressed in this dissertation and additionally references more recent work published in the fields of genome organization and stem cell biology that has been published following publication of Chapter 1.

Chapter 3, the discussion, places the individual findings presented here in the context of the genome organization field as a whole, discusses consequences of a functionally segregated genome, and discusses future directions of this work.

Chapter 4 describes experimental and analytical methods.

CHAPTER 1

Introduction: Pluripotency in 3D: Genome organization in pluripotent cells

1.1 - Abstract

Cells face the challenge of storing two meters of DNA in the three-dimensional (3D) space of the nucleus that spans only a few microns. The nuclear organization that is required to overcome this challenge must allow for the accessibility of the gene regulatory machinery to the DNA and, in the case of embryonic stem cells (ESCs), for the transcriptional and epigenetic changes that accompany differentiation. Recent technological advances have allowed for the mapping of genome organization at an unprecedented resolution and scale. These breakthroughs have lead to a deluge of new data, and a sophisticated understanding of the relationship between gene regulation and 3D genome organization is beginning to form. In this review we summarize some of the recent findings illuminating the 3D structure of the eukaryotic genome, as well as the relationship between genome topology and function from the level of whole chromosomes to enhancer-promoter loops with a focus on features affecting genome organization in ESCs and changes in nuclear organization during differentiation.

1.2 - Introduction

Embryonic stem cells (ESCs), isolated from the inner cell mass of pre-implantation blastocysts, self-renew indefinitely under appropriate culture conditions and have the ability to produce cell types from all three germ layers upon induction of differentiation *in vivo* and *in vitro*[6,7]. Linear genomic features, such as the location of transcription factors, the basic transcriptional machinery, and chromatin modifications, as well as DNase hypersensitivity, expression state, and replication timing have been extensively mapped in ESCs. Therefore, gene regulatory processes controlling the transcriptional program of ESCs are relatively well characterized (reviewed recently elsewhere[8]) and center on three core transcriptional networks: the

pluripotency network, made up of highly expressed, ESC-specific genes bound by the transcription factors Oct4, Sox2, and Nanog which, together, control pluripotency through co-binding of many enhancers and promoters including their own[8]; the cMyc network, formed by transcription factors of the Myc family which drives gene expression by promoting the release of paused polymerase at its target genes[9]; and, the Polycomb group (PcG) protein network, which represses developmental and lineage specific genes[10] through the tri-methylation of lysine 27 of histone 3 (H3K27me3)[11], H2AK119 ubiquitylation[12], and chromatin compaction[13]. These transcriptional networks work in concert with external signaling pathways to maintain the pluripotent state, most notably the LIF-Jak-Stat pathway in mouse ESCs[14] and bFGF-signaling in human ESCs[15]. Highlighting the importance of these transcriptional networks to pluripotent cell identity, ectopic expression of Oct4, Sox2, cMyc, and the pluripotency-associated transcription factor Klf4 is sufficient to reprogram somatic cells to induced pluripotent stem cells (iPSCs)[16]. iPSCs carry all the typical characteristics of ESCs including self-renewal, expression of the endogenous pluripotency program, and differentiation in both the teratoma and chimera formation assays[16]. More recently, there has been a push towards determining genome organization and correlating 3D topology with genomic functions such as transcriptional regulation. Because transcriptional networks and gene regulation are well studied in ESCs, these cells are a great model system with which to understand 3D genome organization and its changes upon cell fate change. In this review, we will first summarize general aspects of genome organization revealed from work with various cell types and then focus on new findings that begin to address genome organization in ESCs and changes upon induction of differentiation.

1.3 - Widely conserved features of genome organization: A top down view

Years of research from many groups utilizing a variety of cell types from numerous species have defined a number of general features of eukaryotic genome organization. Interphase chromosomes reside in discrete, minimally overlapping chromosome territories (CTs, reviewed exhaustively by the Cremer brothers[17], Figure 1.1a). CTs are organized such that small, gene rich chromosomes tend to pair and localize to the nuclear interior[3,18,19]. Cell type-specific radial positioning of CTs within the nucleus has also been reported[20], although the extent to which CT pairing and positioning are conserved through mitosis varies depending on the cell type analyzed[21,22]. Individual genes are largely confined to their respective chromosome's territory, however, in certain developmental contexts, such as *Hox* gene activation[23] and X-chromosome inactivation[24] (discussed in more detail below), gene loci have been shown to loop out or move to the outer edges of their CTs.

Localization of genomic regions to the nuclear periphery, specifically the nuclear lamina, is correlated with gene silencing across the eukaryotic kingdom[25-27], and ectopic targeting of genetic loci to the nuclear envelope (NE) can induce transcriptional silencing in some cases[25,28,29]. NE-mediated gene silencing is thought to function in part through the interaction of heterochromatin protein 1 (HP1) with repressive protein complexes localized to the NE through interactions with the B-type lamins, the major constituents of the NE (reviewed extensively elsewhere[30]), as well as through histone-LaminA interactions[31]. Sequestration of the transcriptional machinery away from the nuclear periphery has been suggested as an additional mechanism of NE-mediated transcriptional silencing, although it is unclear if this phenomenon is a general feature of eukaryotic genome organization[32]. Recent work has added a new player in targeting specific genomic regions to the NE, the vertebrate homologue of the *Drosophila*

GAGA factor, cKrox. cKrox binds GA repeat-enriched lamina associating DNA sequences (LASs) in a cell type-specific manner, targeting these regions to the NE, although it is currently unclear how cKrox is targeted to specific LASs[33].

Early studies of genome organization relied on cytological methods such as fluorescence *in situ* hybridization (FISH), and as such were limited in the number of gene loci that could be analyzed in a single experiment. The past decade has witnessed the introduction of molecular techniques and high-throughput mapping to the field of genome organization in the form of chromosome conformation capture (3C)-based techniques. 3C allows for a molecular view of genome organization via chemical fixation, restriction enzyme digestion, ligation of juxtaposed DNA fragments and detection of ligation events by PCR. The juxtaposition frequency of two DNA fragments in 3D space can be inferred based on the quantity of the PCR product produced upon amplifying a given ligation event[34]. In recent years a number of groups have expanded 3C-based molecular techniques[35] to include 4C – which allows for the identification of all chromatin contacts made by a single locus with the rest of the genome [36,37], 5C – enabling the identification of all pair-wise chromatin interactions for a given genomic region[38], Hi-C[3] and its technical variants[4,39,40] – permitting the identification of all pairwise chromatin interactions genome-wide, and ChIA-PET[41] – allowing the identification of all pairwise chromatin interactions genome-wide, which share binding of a protein of interest.

These techniques have revealed a previously unappreciated hierarchical organization of eukaryotic genomes. As expected from the CT-based structure of the genome, intra-chromosomal (*cis*) chromatin interactions mapped by 3C-based techniques are much more

frequent than inter-chromosomal (*trans*) ones[3,36]. Apart from verifying the existence of chromosome territories and the preferential pairing of small, gene rich chromosomes, mapping of genome-wide chromatin interactions with Hi-C in human lymphoblasts[3], mouse pro-B cells[2], and *Drosophila* embryos[4] demonstrated the existence of a further organizational sub-division of the genome into 'A' and 'B' compartments, where the A compartment is enriched for features of euchromatin and the B compartment is depleted of these features[3]. From an organizational standpoint, chromatin interactions within compartments are much more frequent than those between compartments (Figure 1.1b).

The comparatively smaller size of the *Drosophila* genome allowed for higher resolution DNA topology mapping than was previously accomplished in mammalian genomes and led to the identification of a further organizational sub-division of the genome into linear domains with shared epigenetic features, ranging in size from 10 kilobases (kb) to 500kb[4]. These domains appear to act modularly in governing global genome organization in *Drosophila*. Interactions of loci within a given domain are more frequent than interactions between loci in different domains. However, where inter-domain interactions occur, active domains preferentially interact with other active, domains, inactive with inactive, and PcG-regulated with other domains of PcG enrichment[4]. Recent work with a number of different cell lines has identified analogous domains in mammalian genomes[42-44], termed topological domains or topologically associating domains (TADs). TADs delimit the range within which enhancers can affect their target genes, as co-regulated enhancer-promoter groups tend to form extended clusters of interacting chromatin that align with TADs[45] (Figure 1.1c). Additionally, the changes in gene expression upon differentiation are more likely to occur in the same direction for genes within a TAD than for

genes in different TADs[43]. It has long been appreciated that enhancer-promoter interactions are responsible for regulating the cell type-specific expression of genes. The importance of looping between promoter and enhancers for gene regulation is highlighted by data from the ENCODE consortium showing that genes whose transcriptional start sites are contacted by an enhancer are more highly transcribed than those that are not[46].

The locations of TAD boundaries are strongly conserved between the mouse and human genomes, particularly within syntenic regions; and TADs of both species are largely conserved across different cell types[42,43]. CP190, a critical contributor to the function of various *Drosophila* insulator proteins through its mediation of DNA looping[47], is enriched at TAD boundaries in *Drosophila*, and the vertebrate insulator protein CTCF[48] is similarly enriched at the boundaries of a large subset of mammalian TADs[42,44], suggesting an evolutionarily conserved mechanism of TAD formation by insulator proteins, similar to what has been proposed for mammalian insulators in general[49]. In ESCs, CTCF has been shown to mediate DNA looping events which partition the genome into physical domains each characterized by distinct epigenetic states[44], supporting a model of DNA organization wherein many TADs function as large, independently regulated DNA loops (Figure 1.1b,c). Although the data arguing for the role of insulator proteins in delimiting TAD boundaries is strong, it is worth noting that only a portion of insulator binding sites function as TAD boundaries in mammalian and *Drosophila* cells[4,42], and that many enhancer-promoter interactions cross CTCF binding events in a variety of mammalian cells types[46]. More work will therefore be required to determine the necessary and sufficient constituents of TAD boundary delimiters.

Albeit in flies the interactions of TADs has been described[4] (see above), the extent to which mammalian TADs interact with each other, and the mechanistic logic behind these interactions, remains unclear. Distal chromatin interactions between loci many millions of bases (Mb) apart, or in *trans*, have been demonstrated in a number of mammalian cell types by various 3C-based studies[3,36,37,44,46], but these interactions have not been examined in the context of TADs. It has been shown that long-range chromatin contacts can be cell type-specific and can occur between regions of the genome enriched for the DNA binding motif of a given transcription factor or for genes regulated by the same trans acting factors[50,51], or by binding of gene regulatory factors as has been demonstrated for PcG-regulated distal chromatin interactions in *Drosophila*[4,52] (Figure 1.1). One may speculate that co-regulated TADs are brought together in physical space in mammalian genomes as a general rule[4]. Comprehensive analysis of long-range interactions in a well-annotated cell type such as ESCs should contribute to a better understanding of this question, as gene regulatory networks are well understood[8], and - in the case of mouse ESCs - are amenable to genetic manipulations, which can be used to test causal links between linear genomic features and genome organization both in pluripotency and during the course of differentiation.

1.4 - The ESC genome in pluripotency and differentiation

The genomes of ESCs have a number of unique characteristics that distinguish them from somatic cell genomes. The contribution of these features to the different layers genome organization described above is currently unclear, however they may have an effect on the interpretation of organizational data in ESCs and thus are important to note. Among features unique to the genome of mouse ESCs are a hyper-dynamic association of chromatin proteins

with the chromatin polymer[53], enhanced global transcriptional activity[54], a lack of condensed heterochromatin at the NE and peri-nucleolar regions[55], and two active X-chromosomes in females cells. Upon differentiation, chromatin protein association becomes more stable[53], wide-spread transcription of both protein coding and non-coding regions is restricted, repeat elements are silenced[53,54], and heterochromatic regions of the genome compact and localize to the nuclear periphery[55]. At the same time, a subset of pluripotency gene loci is silenced and moves to the nuclear periphery even before germ layer restriction occurs[55-58]. These processes occur contemporaneously with large-scale changes in DNA replication timing[56,57], silencing of an X-chromosome in female cells[59], and the onset of LaminA expression, which stabilizes histone H1 in heterochromatin and is required for the establishment of the large number of heterochromatin foci characteristic of differentiated cells[60]. Together, these data indicate that the dramatic changes in gene expression that occur upon pluripotent cell differentiation are accompanied by large-scale changes in genome topology.

Despite the correlation between NE localization and gene silencing in ESCs[58], LaminB1/B2 double knockout ESCs and trophectoderm cells show few changes in gene expression compared to their respective wild-type cells, and those genes that do change expression levels are not bound by B-type Lamins in wild-type cells[61]. This suggests that LaminB does not directly regulate expression of its interacting genes in ESCs or trophectoderm cells. Alternatively, unidentified redundant mechanisms may work to maintain gene silencing at the NE in the absence of B-type lamins in these cells. Additionally, LaminB-null ESCs show none of the NE morphology defects typical of somatic cells with mutations in nuclear lamina proteins[61,62]. During the course of differentiation of ESCs to neural precursor cells, many

pluripotency specific genes are re-localized to the nuclear lamina and many NPC-specific genes detach from the lamina[58]. In contrast to the phenotypically wild-type ESCs, upon embryonic development, LaminB1/B2-null mice display severe organogenesis and neural migration defects[61]. Implicated as a major player in somatic cell genome organization, it will be important to understand the role of the nuclear lamina in regulating genome organization of ESCs, or alternatively, to determine if chromatin-NE co-localization is only required upon differentiation.

In contrast to the transcriptionally repressive nuclear envelope, in yeast, gene localization to the nuclear pore complex is associated with transcriptional activation in certain inducible systems[63]. In metazoans, however, some of the nucleoporins (Nups), the major constituents of the nuclear pore complex, have been implicated as regulators of gene expression through direct binding of chromatin in the nucleoplasm, mostly away from the nuclear pore[64-66]. Specifically, Nup133-null mice display defects in neural differentiation and Nup133-null ESCs differentiate inefficiently along neural lineages and do not contribute to the neural tube of chimeric embryos[67]. Similarly, the integral membrane protein Nup210 is expressed cell type-specifically and is not essential for nuclear pore function, but is required for ESC differentiation into neural progenitors as well as for myogenesis. Nup210 depletion abrogates the upregulation of differentiation-associated genes and its overexpression facilitates the expression of essential differentiation genes. Notably, the authors argue against a role for Nup210 in tethering genes to the nuclear pore complex upon induction, as they do not see changes in candidate, Nup210 regulated gene localization to the NE[68]. It will be important to understand the differing roles of Nups when they are chromatin bound in the nucleoplasm, versus when they are part of the nuclear pore complex, as well as their role in genome organization or re-organization upon differentiation in metazoans.

1.5 - Re-organization of the X chromosome during ESC differentiation

The X chromosome inactivation process is a striking example for topology changes associated with differentiation. The equalization of X-linked gene expression between sexes in mammals occurs via the silencing of one of two X chromosomes upon induction of differentiation of ESCs. This process, induced by the up-regulation and spreading of the non-coding RNA *Xist* on the future inactive X chromosome (Xi), leads to the transcriptional silencing of the majority of X-linked genes on the Xi, and the establishment of a number of repressive chromatin modifications along the Xi, including Polycomb group protein-mediated H3K27 methylation, DNA methylation, and deposition of the histone variant macroH2A[59].

At the onset of X-inactivation homologous X chromosomes co-localize allowing for the pairing of the *Xist*-encoding X-inactivation centers (XIC), a process thought to be necessary for the initiation of X-inactivation on one of the two X chromosomes[69-71]. Following X-chromosome pairing, the Xi preferentially localizes to the NE and peri-nucleolar regions of the nucleus[72], both of which are enriched for autosomal heterochromatin in differentiated cells[73]. This localization occurs predominantly during S phase of the cell cycle, and is dependent on *Xist* expression. Deletion of *Xist* in fibroblasts causes a re-localization of the Xi away from the nucleolus, with concomitant re-activation of a subset of genes in a small proportion of cells[72].

In addition to these large scale movements of the X-chromosome upon induction of X-inactivation, *Xist* expression leads to the formation of an *Xist* RNA domain over one of the two X-chromosomes and the immediate exclusion of RNA polymerase II (RNAPII) and transcription machinery from the future Xi[24]. Interestingly, the exclusion of transcription machinery precedes

the completion of transcriptional silencing. At the time of transcription machinery exclusion from the territory of the Xi, actively transcribed genes localize to the periphery of the X chromosome territory, where they contact the transcriptional machinery. As these genes are silenced during the course of differentiation and X-inactivation, they localize to the interior of the Xi territory. Silencing and sequestration of X-linked genes into the Xi territory requires the A-repeat[24], a portion of *Xist* necessary for transcriptional silencing[74]. Genes that escape X-inactivation remain localized to the periphery of the Xi territory[24]. A subsequent 4C study has shown that these escaping genes co-localize with other escaping genes as well as with gene loci on other chromosomes[75]. Conversely, silenced genes in the center of the Xi territory make few preferential interactions with other genomic regions, suggesting a random localization or restricted movement of these loci within the Xi[75]. Xi-specific 3D chromatin organization is partially dependent on *Xist* RNA coating, as *Xist* deletion results in an organizational state of the Xi resembling the Xa[24,75].

The mechanisms regulating the dramatic re-organization of the X-chromosome upon silencing are unclear, however, SatB1/B2 are implicated in this process[76]. In thymocytes, the SatB1 protein is organized in a cage-like structure throughout the nucleus[77] where it regulates gene expression through the anchoring of looped chromatin structures and the recruitment of chromatin modifying enzymes[78,79]. Upon induction of *Xist* expression in thymocytes and ESCs, *Xist* RNA accumulates in a region delimited by SatB1, and SatB1 depletion during ESC differentiation reduces the efficiency of X-inactivation[76], although MEFs derived from SatB1/B2-null embryos display normal X-inactivation[80,81], calling into question an essential role for SatB1 in the organization of chromatin and gene silencing during X-inactivation.

Despite the large-scale re-organization of the X chromosome during the course of inactivation, the existence of the two TADs encompassing the XIC does not change. However, specific intra-TAD interactions are lost upon X-inactivation, suggesting a random organization of the intra-TAD space within the Xi[43], similar to that shown for long-range interactions with in the Xi by 4C analysis[75]. Alternatively, molecular 'gluing' of these TADs to the nuclear lamina could lead to a very limited interactome. Cell lines lacking G9a, an H3K9 methyltransferase, or Eed, an essential component of the Polycomb repressive complex 2, have no effect on the chromatin conformation or TAD structure within the XIC, suggesting that epigenetic modifications function downstream of TAD formation. In contrast, deletion of the TAD boundary region in the XIC, specifically between *Xist* and *Tsix*, resulted in the partial merger of neighboring TADs in mouse ESCs[43], although cells lacking this TAD boundary are still capable of undergoing random X-inactivation upon differentiation[82], leaving open the question of whether a specific organization of the XIC is required for X-inactivation.

Together, these data argue that X-inactivation is an essential developmental process that is associated with topology changes at various levels and may be a great model system to dissect the molecular mechanisms underlying genome organization and its dynamics during the course of differentiation. Notably, the 3D organization of the X-chromosome during Xi-reactivation events *in vitro* or *in vivo*, either in the context of somatic cell reprogramming[83] or germ cell development[84], has not been investigated.

1.6 - Mechanistic insights into genome organization

Based on studies of promoter and enhancer interactions by DNA looping, it is clear that gene expression is facilitated and regulated through distal chromatin contacts. The mode and mechanism of action of enhancer elements has been the subject a large body of work over the years, and recent experiments have brought to light various molecular mechanisms underlying this phenomenon. In particular, the Cohesin complex – which, canonically, forms a ring around sister chromatids during mitosis[85] - has been shown to play a major role in organizing DNA topology and affecting gene regulatory processes at the level of enhancer-promoter interactions. It was initially characterized at the developmentally regulated *IFNG* locus in T-cells where it is required for enhancer-promoter looping and expression of *IFNG*[86], and at the *H19/IF2* loci in humanized mouse cells where it is required for insulator activity[87]. Cohesin binding sites overlap significantly with CTCF binding sites genome wide[87-90], many of which are conserved across cell types and species[49], leading to a model wherein CTCF-associated Cohesin localization is largely cell type invariant[91] (Figure 1.1b,c), potentially explaining the conservation of TAD boundaries across cell types and species, as hypothesized by Dixon et al[42,49].

In order to generate cell-type specific DNA topologies for the facilitation of specific transcriptional programs, cells appear to utilize non-CTCF mediated recruitment of Cohesin. For instance, Cohesin is co-bound with the transcription factor CEBPA in Hep2G cells and with the estrogen receptor (ER) in MCF7 cells, where Cohesin binding persists in the absence of CTCF[92]. In the case of MCF7 cells, Cohesin binding is particularly enriched at regions involved in ER-mediated chromatin interactions[41]. Mounting evidence suggests that, similar to its role during

mitosis, Cohesin functions by holding functional DNA elements together in the nucleus (Figure 1.2), and additionally, may stabilize TF binding to highly occupied *cis* regulatory elements[93].

A major advance in our understanding of the mechanistic underpinnings of promoter-enhancer interactions in ESCs was achieved recently through an shRNA screen for loss of *Oct4* gene expression[91]. This screen identified numerous subunits of Mediator - a massive protein complex that regulates the activity of RNA Polymerase II[94] - and Cohesin subunits, as well as the Cohesin loading factor Nipbl, as regulators of *Oct4* gene expression. The authors found that Cohesin and Mediator co-immunoprecipitate with each other and Nipbl in ESCs, potentially allowing Cohesin to enable ESC-specific enhancer-promoter interactions upon recruitment of Mediator to chromatin by various transcription factors (Figure 1.2). Unlike CTCF and Cohesin co-bound sites, Mediator and Cohesin co-bound sites are cell type-specific and often overlap with locations of pluripotency transcription factors Oct4, Sox2, and Nanog in ESCs. In MEFs, among loci where Mediator binding is different compared to ESCs, enhancer-promoter looping interactions are likewise different, as shown using 3C at a number of candidate loci[91]. These findings likely explain previous work demonstrating a chromatin topology that brings together a variety of DNase HS sites and co-regulated genes within the extended 150kb *Nanog* locus, a topology that is lost upon Oct4 depletion[95]. Although it has not been explicitly demonstrated outside of ESCs, we speculate that recruitment of mediator to binding sites occupied by cell type-specific transcription factors facilitates the recruitment of Cohesin to interphase chromatin where it mediates enhancer-promoter interactions, and potentially even more long-range chromatin contacts.

1.7 - Conclusions and Outlook

The synthesis of recently published data leads us to propose the following speculative model of mammalian genome organization (Figure 1.1): Within TADs[4,42,43], enhancers and promoters dynamically co-localize with and co-regulate each other[46,96] in a cell type-specific manner[46], limited in range along the chromatin polymer by TAD boundaries[43,45]. These TADs, existing as topologically isolated loops[44], can re-locate to various sub-nuclear compartments[3,23,43,58] in response to specific developmental and gene regulatory cues, but apart from limited cases where specific genes (and likely entire TADs) loop out of their CTs, TAD localization is limited to its own CT. An important piece of information missing from this model is the mode and mechanism of preferential TAD-TAD interactions that we infer from 4C data. Due to their well-defined transcriptional networks, chromatin states, and gene expression data sets, ESCs - in pluripotency and during the course of differentiation - will be an ideal cell type for studying this question with 3C-based methodologies. In combination with a transcriptionally permissive nuclear environment[54] and a lack of highly condensed heterochromatin[55], future studies may also help us to understand whether an ESC-specific 3D genomic organization contributes to the developmental plasticity of ESCs.

CHAPTER 2

Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization

2.1 - Summary

The spatial organization of the genome is linked to its function, but the relationship between genome organization and gene-regulatory networks that govern cell identity remains unclear. Here, we mapped long-range chromatin interactions in embryonic stem cells (ESCs), iPSCs, and fibroblasts, and uncovered an ESC-specific genome organization that is gradually re-established during reprogramming. Confirming previous results, we show that open, accessible and closed chromatin character is the primary determinant of long-range chromatin interaction preferences. Importantly, we find that in ESCs, genomic regions extensively occupied by the pluripotency factors Oct4, Sox2, and Nanog preferentially co-localize. Similarly, regions strongly enriched for Polycomb-proteins and H3K27me3 frequently interact, and loss of the Polycomb-protein Eed diminishes these interactions without dramatically changing overall chromosome-conformation. Consistent with a spatial segregation of gene-regulatory networks in ESCs, Nanog and Polycomb-proteins occupy distinct nuclear spaces. Together, these data reveal that transcriptional networks that govern ESC-identity play a role in determining genome-organization.

2.2 - Introduction

The invention[34] and expansion of chromosome conformation capture (3C)-based technologies[97] have led to a new paradigm wherein gene regulation can be studied in the context of three-dimensional (3D) genome organization[98,99]. Recent work has also demonstrated an organizational hierarchy to metazoan genome structure[100]. At the smallest scale, considering up to a few hundred kilobases (kb) of linear DNA, enhancers and promoters come into physical contact to establish cell type- specific expression programs[45,46,101]. These interactions are maintained by the Cohesin complex, which can be recruited to interphase chromatin via the mediator complex[91], as well as by cell-type specific transcription factors[5,92,102]. In mammals, promoter-enhancer interactions are confined to topologically associating domains (TADs), ~1 megabase (Mb) cell type-invariant, self-associating genomic regions whose boundaries are enriched for the insulator protein CTCF[42,43].

As a second level of the organizational hierarchy, TADs are thought to function as the fundamental modular unit of gene regulation and genome organization[42,43], with changes in gene expression and nuclear lamina association that accompany cellular differentiation often occurring in a TAD-wide manner[43,45]. Due to the preferential co-localization of specific TADs, gene regulatory elements, or genes within TADs, long-range chromatin interactions between genomic regions many Mb apart on the same chromosome (*cis* or intrachromosomal) or on different chromosomes (*trans* or interchromosomal) have also been identified[36,37,50,51,103-106], and constitute the third level of the organizational hierarchy. Although such spatially co-localizing distal genomic regions have been reported to be enriched for the sequence motifs of specific transcription factors[51], co-expressed genes[103], or co-regulated genes[50,104,106], insight into the

relationship between long-range chromatin interactions and the regulatory features enriched in the col-localizing regions is still limited.

In this study, we explore long-range chromatin contacts in mouse ESCs using 4C-seq in an effort to understand the relationship between chromatin contacts and the gene regulatory factors that govern cell identity. We define the distal chromatin interactions made by a variety of genomic 'bait' regions, chosen to represent different chromatin states[10,107-109] and pluripotency-specific transcription factor binding profiles[110-114]. We further examine how genome organization changes in the absence of a critical chromatin regulator, and upon differentiation and transcription factor-induced reprogramming of differentiated cells to iPSCs. Together, our data suggest a previously unappreciated hierarchy in the organization of long-range chromatin contacts and reveal that distal genomic regions that share common gene regulatory characteristics co-localize within the three-dimensional space of the nucleus.

2.3 - Results

2.3.1 - Experimental approach to studying chromatin contacts

To investigate the distal chromatin interactions occurring between genomic regions mega-bases away on the same or different chromosomes in ESCs, we performed chromosome conformation capture (3C)[34] coupled to high throughput sequencing (4C-seq)[1] for 16 bait regions (Table 1.1). 4C-seq allows one to the identify any mappable genomic region in close physical proximity to a specific genomic (bait) locus within a population of cells at the moment of fixation by means of proximity-ligation of juxtaposed DNA fragments. The product is a library of chimeric DNA fragments containing the bait region and its interacting DNA partner(s) ligated at a restriction site, in our case HindIII, that can be identified by high-throughput sequencing. The resulting 4C-seq reads were aligned to unique HindIII sites, which we defined to be those 'AAGCTT' hexamers flanked by unique 50bp sequences on both sides (see *Experimental Procedures*).

Our 4C-seq data were highly reproducible across biological and technical replicates utilizing distinct primer pairs for amplification of ligated fragments and different HindIII fragments within bait regions as anchor point (Figures 2.1A, 2.2A, Table 2.2). As a result, replicate data sets for each bait locus that passed stringent quality control requirements were pooled for downstream analysis by calculating the probability that each HindIII site recorded a read across all replicates (Figure 2.1D/E, Table 2.3, see *Experimental Procedures*). To obtain a semi-quantitative measure of interactions, we calculated the average hit probability within 200kb windows along the genome, which we refer to as the 'hit percentage' (see *Experimental Procedures*). For all of our baits, the hit percentage was higher in *cis* than in *trans,* as exemplified by the 4C-seq interactome of the *Pou5f1*-encoding region (Figure 2.3A).

Furthermore, as expected for 3C-based techniques and the nature of chromosome conformation[3], intrachromosomal interactions decay rapidly as a function of genomic distance to the bait along the linear DNA (Figure 2.3Aii).

Significantly interacting regions were identified by a binomial test as those 200kb windows, centered on HindIII sites, that showed a markedly higher hit probability than expected (Figures 2.3A/B, 2.4A, Table 2.1). In *cis*, we empirically modeled the expected background hit probability by averaging the hit probability profiles across all our ESC data sets as a function of distance from the respective bait loci (Figure 2.3A,iii) (see *Experimental Procedures)*. This model accounts for the preference for bait-proximal ligations, thereby allowing us to identify those distal regions that interact more frequently than expected based on genomic distance to the bait. In *trans*, the average hit probability across each chromosome was used as the expected background level of interaction, since interchromosomal interactions show no detectable positional biases with respect to the bait locus. Simulated data, sampled from the empirically modeled expected hit probability, were used to estimate false discovery rates (Table 2.4**,** see *Experimental Procedures*). To test the reliability of our analysis and 4C-seq data, we partitioned our *Pou5f1* 4C-seq replicates into two equal subsets, pooled the libraries within each subset, and ran the pooled, partitioned data through our 4C-seq analysis pipeline. We obtained highly overlapping intrachromosomal interactomes from the partitioned data sets (Figure 2.1B). As expected, given the inherent bias for interactions to occur in *cis*[3], *trans* interactions are less frequent and therefore less reproducible, and exhibited higher FDRs, but nonetheless showed highly significant overlap between the partitioned data sets (Figure 2.1B).

To confirm the distal chromatin interactions identified by our 4C-seq method, we utilized several approaches: First, the presence of various ligation products between distal regions was corroborated using 3C[115] (Figure 2.1F). Second, we performed three-dimensional (3D) fluorescence *in*-situ hybridization (FISH) and showed that an interacting genomic region 52Mb away from *Pou5f1* bait region on the same chromosome (chr17) was significantly closer to the *Pou5f1* locus in 3D-space than a non-interacting locus located in an intervening region only 35Mb away (Figure 2.3C). Third, two 4C-seq experiments in which intrachromosomal interaction sites of the *Pou5f1* region were used as baits demonstrated the reciprocation of *Pou5f1* 4C-seq defined chromatin interactions (Figure 2.3D). Fourth, control 4C libraries prepared from either unfixed cells or from genomic DNA showed no correlation with experimental data sets (data not shown), and produced no significant interactions (Figures 2.1A/C). Finally, our 4C-seq interactomes correlated strongly with data from a previously published genome-wide interactome based on Hi-C[42] (Table 2.5, Figure 2.6H), even though 4C-seq sequencing depth is much greater than that allowed by Hi-C for a given bait region. Together, these findings validate our 4C-seq defined chromatin interactions and rule out technical biases that can affect data generated by 3C-based studies (Figure 2.1G).

Noteworthy, the analysis of the *Pou5f1* 4C-seq data revealed an extensive intrachromosomal interaction network, raising the question of how the spatial interactions of this important locus are regulated in ESCs (see below).

2.3.2 - A pluripotency-specific organization of the mouse genome

As a first step towards understanding the regulation of distal chromatin interactions in ESCs, we determined whether the spatial interactions of the *Pou5f1* locus change upon differentiation. Because the *Pou5f1* gene encodes the transcription factor Oct4 that is essential for the establishment and maintenance of ESCs and is specifically expressed in the pluripotent state[116], we determined whether the changes in expression and chromatin state that are associated with the silencing of the *Pou5f1* locus during differentiation[117] coincide with changes to that locus' interactions in 3D space. We found that the *Pou5f1* interactome in mouse embryonic fibroblasts (MEFs) is distinct from that of ESCs, both in terms of hit probability (Figures 2.5A, 2.2A) and interacting domains (Figure 2.5B) across the *cis* chromosome. To obtain additional interaction perspectives in both cell types, we determined the interactomes of nine more baits in MEFs for which we already had generated their ESC profiles (Table 2.1, Figure 2.2A). Their interactomes also differed significantly between ESCs and MEFs (Figures 2.5E/F). For some baits, such as the *Dppa2* region, the chromatin contacts are more dramatically different between MEFs and ESCs than those of the *Pou5f1* region (Figures 2.5C/D, 2.2A). Notably, the *Dppa2* region has been shown to reposition towards the nuclear periphery and switch interaction preferences in *cis* from regions of early DNA replication to regions of late DNA replication during ESC differentiation[56,118], consistent with a major change in distal chromatin contacts between pluripotent and differentiated cells genome-wide detected by our 4C-seq experiments. These data highlight an ESC-specific organization of the genome, and suggest that changes in 3D-interactions during differentiation are regulated at a locus-specific level.

We next tested whether the 3D-organization of the MEF genome can be reset to an ESC-like state by transcription factor-mediated reprogramming to induced pluripotent stem cells (iPSCs)[16]. To this end, we determined the interactomes of eight of our bait regions including the *Pou5f1* locus for two additional cell types that represent distinct stages of the reprogramming process (Table 2.1). First, we performed 4C-seq on faithfully reprogrammed iPSCs that have been derived from MEFs, and found that the distal chromatin contacts in iPSCs are highly similar to those of ESCs, both in terms of hit probability across the *cis* chromosome (Figure 2.5A/C/E, 2.2A), and interacting domains in *cis* and in *trans* to the bait region (Figures 2.5B/D/F; 2.2B). Second, we examined chromatin interactions in pre-iPSCs, a clonal population of cells that represents a late reprogramming stage at which the pluripotency expression program is not yet induced[111]. We found that the interactomes in these cells are typically distinct from those in pluripotent cells (ESCs and iPSCs) as well as those in MEFs (Figures 2.5, 2.2), suggesting that, similar to transcriptional and chromatin states[111], at a late stage of reprogramming the re-organization of chromatin contacts is not yet complete, in line previous findings detailing the long-range interactions of the *Nanog* locus[119]. Together, these data show that the large-scale changes in genome organization that occur during differentiation are gradually reset to an ESC-specific state during reprogramming to the iPSC-state, likely associated with progressive changes in chromatin and transcription states[111,120,121]. The cell type-specificity of long-range chromatin interactions contrasts with the largely cell-type invariant TAD structure of the mouse[42,43], human[42], and fly[4] genomes, suggesting that the organization of long-range chromatin contacts is a key feature of cell identity.

2.3.3 - Open and closed chromatin as the foundation of genome organization in ESCs

The finding that changes in distal chromatin contacts occur during differentiation and reprogramming suggested a link between genomic features that function to establish and maintain cell identities, such as chromatin states and transcription factor binding, and genome organization. To investigate this association in the context of ESCs, a cell type for which an abundant array of genomic feature data is available, we compiled a compendium of genomic features in ESCs. Specifically, we considered the binding profiles of the basic transcriptional machinery (TBP, RNA polymerase II), co-activators (Mediator and p300), and architectural proteins (Cohesin and CTCF) (Table 2.6). In addition, we made a point to include transcriptional regulators known to belong to important gene regulatory networks, such as the pluripotency transcription factor network, comprised of Oct4, Sox2, Nanog, and Klf4[110,112-114], the transcription factors that cooperate with cMyc (cMyc, Max, E2F4)[122], and the repressive Polycomb protein network, captured by Ring1b and Polycomb repressive complex 2 (PRC2) occupancy[10] (Table 2.6). Furthermore, since regulatory genomic regions such as enhancers and promoters typically exhibit extensive co-occupancy by these factors, we grouped them into 11 clusters based on the co-binding that occurs between these transcription factors and regulatory proteins at 1kb resolution (Figure 2.6A, see *Experimental Procedures*). We also took into account the combinatorial nature of histone modifications by summarizing the relationship of histone marks in terms of five functionally distinct chromatin states (ChromHMM states)[123]. These were associated with Polycomb repression, transcriptional elongation, enhancers, and promoters, and an additional 'low signal' state was assigned to regions of the genome lacking the chromatin modifications we considered (Figure 2.6B**,** see *Experimental Procedures*). In addition,

replication timing[56], DNase hypersensitivity, gene density (in terms of transcriptional start sites (TSS)), RNA-seq expression data, as well as LaminB association[44] were considered (Table 2.6).

Principal components analysis (PCA) was used to aggregate the resulting 22 linear genomic data sets and reduce their high dimensionality (see *Experimental Procedures*). For downstream analyses, we retained the first three principal components (PCs) (Figure 2.6C). To explore the relationship between genomic features and genome organization in ESCs, we then compared the linear genomic character represented by the principal components to the 4C interactome data for each of the 16 bait regions analyzed in ESCs (Figure 2.4B).

The first principal component (PC1) captures 51% of the variance across all features (Figure 2.6C) and distinguishes open, accessible chromatin and closed, inaccessible chromatin (Figure 2.6D). Specifically, regions of the genome with positive PC1 scores are characterized by high gene density, DNase hypersensitivity, binding of Cohesin, the basic transcriptional machinery, and transcription factors, as well as active or Polycomb-repressed chromatin states. In contrast, regions of the genome with negative PC1 scores are strongly depleted for these features and typically replicate their DNA late in S phase and are LaminB-associated (Figure 2.6D). Notably, the 1Mb regions surrounding each bait locus had widely different PC1 characteristics (Figure 2.6E,i - top panel). Additionally, we found that the mean PC1 scores for the 1Mb bait regions correlated strongly with the mean PC1 scores within the corresponding interactomes, in both *cis* and *trans* (Figure 3E,i, Spearman rho = 0.75 and 0.85, respectively). That is, baits with highly open and accessible chromatin preferentially interact with regions of the genome with similarly high PC1 scores. By contrast, the interactomes of baits within closed

chromatin, i.e. with negative PC1 scores, preferentially co-localize with genomic regions of similarly low PC1 scores in *cis* and *trans*. These findings were confirmed by repeating the 4C-seq analysis for a partially overlapping set of bait regions in an independent ESC line (Figure 2.6E,ii).

To test whether this trend extended genome-wide, we analyzed a previously published genome-wide chromosome conformation capture (Hi-C) dataset for mouse ESCs[42] in a similar manner (see *Experimental Procedures*). For the Hi-C-based analysis, each 200kb region of the genome was treated as a "pseudo-bait" and its intrachromosomal interactome was extracted from the chromosome-wide contact matrix as a row of re-binned, normalized read counts (Figure 2.6F). Genome-wide, we found a striking positive correlation between the mean PC1 score of the 1Mb pseudo-bait regions and their most frequently interacting intrachromosomal sites (Figure 2.6G), corroborating our 4C-seq-based findings. A comparison of the profiles of the most and least likely interacting regions further demonstrated that genomic regions with very open, high PC1 chromatin character, e.g. the *Pou5f1* region, interact extensively with most other genomic regions with similarly high PC1 scores on the chromosome in *cis* and do not interact with genomic regions of negative PC1 scores (Figures 2.6G, 2.4C,i). In contrast, genomic regions with LaminB association and of closed chromatin state, defined by negative PC1 scores, avoid regions of high PC1 character and appear to interact more selectively with only a subset of PC1 negative genomic regions (Figures 2.4C,i). This may reflect a tethering of these negative PC1 genomic regions to the nuclear lamina, thus limiting their sampling of distal interactions. The positive correlation between bait and interactome character in terms of their PC1 scores persisted when only distal chromatin contacts (further than 10 Mb away from the bait region) were

considered (Figure 2.4D,i), demonstrating that local contacts close to the bait locus do not drive the correlation, and that interacting regions far away from the bait in *cis* also show an association with chromatin of similar open/closed character.

Taken together, we conclude that genomic loci with a similar PC1 character preferentially interact (or co-localize) within the 3D-space of the ESC nucleus, both in *cis* and in *trans*. The Hi-C-based results indicate that these associations are a general feature of long-range chromatin interactions. Our findings suggest that the strong interaction preferences between regions of similar PC1 character appear to be an intrinsic aspect of overall chromosome conformation, in line with previous findings[3,36,124]. Furthermore, the data demonstrate that interaction preferences in ESCs are not accurately described by binary model of open versus closed chromatin states (i.e. genomic regions with an open chromatin character co-localize with any other open chromatin region and vice versa)[3]. Instead, the ESC interactome follows a more graduated model, where high PC1 regions predominantly interact with regions of similarly high PC1 scores, mid PC1 regions with other mid to low PC1 regions, and negative PC1 regions predominantly with other negative regions, supporting findings in other cell types[124]. In agreement with this, PC1 scores across the ESC genome are continuous rather than discrete (Figure 2.4B,i), and correlate with the continuum of ESC contact frequencies.

2.3.4 - 3D-interactions between Oct4/Sox2/Nanog- and Polycomb-enriched genomic regions in ESCs

Next, we wanted to explore the extent to which long-range interactions are associated with cell type-specific transcriptional networks, beyond their association with the large-scale open versus

closed chromatin properties demonstrated by the correlation of interactomes with the genomic PC1 character. Since PC1-positive regions reflect enrichment for multiple features with diverse functionalities, we reasoned that an interactome's PC1 character may not necessarily reflect any specific mechanistic role of these features, but is most likely a consequence of the overarching chromosomal conformation framework (Figure 2.7A-D). Therefore, we considered the second and third principal components (PC2 and PC3), which account for seven and six percent of the variance in genomics features, respectively (Figure 2.6C). Importantly, these components capture regions of the genome particularly enriched for previously described gene regulatory networks in ESCs (Figure 2.8A/E). Specifically, positive PC2 scores mark regions of exceptional Mediator and Cohesin binding enrichment, as captured by transcription factor (TF) clusters 5, 6, and 11; while negative PC2 regions are enriched for binding of transcription factors belonging to the pluripotency network including Nanog, Oct4, and Sox2 as well as p300, which co-localize as TF clusters 7 and 9 (Figure 2.8A). Figure 2.8B displays the relationship between PC1 and PC2 in terms of these features: whereas genomic regions of open chromatin (defined by positive values of PC1) generally have a higher level of pluripotency factor binding than those with negative PC1 scores, genomic regions with negative PC2 scores have an increased density of pluripotency factor co-occupancy over and beyond what is explained by their chromatin's accessibility (i.e. PC1 state) alone.

Conversely, regions of the genome with positive PC3 values are highly enriched for occupancy by the polycomb protein complexes PRC1 and PRC2 as well as H3K27me3. Negative PC3 regions, on the other hand, seemingly capture the recently described super-enhancers associated with highly transcribed genes in pluripotent cells[125], and represent a strong enrichment

for active transcriptional elongation along with dense occupation by Mediator, Cohesin, the cMyc complex, and the pluripotency transcription factors Oct4, Sox2, and Nanog (Figures 2.8E, 2.7E/F). Notably, genomic regions with positive or negative PC3 scores have an elevated density of their characteristic features above what is explained by their chromatin's PC1 scores in general (Figure 2.8F, and data not shown), and are enriched for genes implicated in developmental processes such as homeobox genes and chromatin organization, respectively, based on gene ontology analysis. We conclude that PC2 and PC3 describe a finer level of chromatin structure associated with specific transcriptional networks that is not captured by the open and closed chromatin character defined by PC1.

To examine whether the genomic states captured by PC2 and PC3 are associated with long-range chromatin contacts, we examined the PC2 and PC3 scores within our 16 ESC bait regions and their respective interactomes in an analogous manner to our analysis of PC1 shown in Figure 3. Importantly, the bait regions exhibited widely different PC2 character in ESCs (Figure 2.8C,i). Furthermore, the mean PC2 score within each baits' intrachromosomal interacting domains showed a strong concordance with the mean PC2 score within the 1Mb bait region itself (Figure 2.8C,i). This same observation was confirmed for additional bait regions in an independent ESC line (Figure 2.8C,ii) and, remarkably, also for *cis*-interactions genome-wide based on Hi-C data analysis (Figures 2.8D, 2.4C,ii/D,ii). The data based on PC2 demonstrate that genomic regions with particularly strong Mediator and Cohesin binding preferentially co-localize in *cis*, as do regions enriched for Oct4, Sox2, and Nanog binding (captured by TF clusters 7 and 9). Notably, however, the co-localization of pluripotency factor enriched, PC2-negative genomic regions, was not evident for *trans* interactions (Figure 2.8C).

The comparison of bait and interactome PC3 scores demonstrated a preferential co-localization of the Polycomb/H3K27me3 enriched, PC3-positive genomic regions (as captured by TF clusters 6, 8, and 10, and the Polycomb Repression ChromHMM state), both in *cis* and in *trans* (Figure 2.8G). For example, *Hox* loci, which belong to the most strongly Polycomb-occupied and H3K27me3-enriched PC3-positive regions in the ESC genome, were found to interact with other regions characterized by high H3K27me3/Polycomb occupancy in *cis* and *trans* (Figure 2.8G). The co-localization of genomic regions occupied by Polycomb proteins and H3K27me3 may be analogous to the co-localization of Polycomb response elements to Polycomb bodies in *Drosophila*[4,52,126,127], and may therefore represent an evolutionarily conserved feature of genome organization linked to gene regulation. Conversely, bait loci with negative PC3 values that are strongly enriched for ESC super enhancers also co-localize with similarly PC3 negative genomic regions both in *cis* and *trans* (Figure 2.8G). Notably, these trends were confirmed and extended genome-wide based on Hi-C data analysis (Figure 2.8H**,** 2.4C,iii/D,iii), showing them to be a genome-wide phenomena.

Together, these results argue that regions of the genome enriched for specific gene regulatory features preferentially co-localize within the 3D space of the nucleus, and raise the possibility that ESC-specific transcriptional and chromatin-regulatory networks are involved in mediating long-range chromatin contacts in ESCs.

2.3.5 - Spatial segregation of Nanog and H3K27me3 in the ESC nucleus

The preferential interactions of genomic regions with positive and negative PC3 scores, respectively (i.e. Polycomb protein versus super enhancer/Oct4/Sox2/Nanog enriched regions), suggested a spatial segregation of genomic regions with opposing PC3 character into distinct compartments in the nucleus. To test this hypothesis, we examined the co-localization of Nanog (PC3 negative), RNA polymerase II (PC3 negative), and H3K27me3 (PC3 positive) in the ESC nucleus by immunostaining. Image analysis showed that Nanog and RNA polymerase II, respectively, have a distinct localization pattern from H3K27me3, with sites strongly enriched for H3K27me3 displaying weak Nanog and RNA polymerase II accumulation and vice versa (Figure 2.9A,B). By contrast, RNA Polymerase II and Nanog overlap more extensively, albeit not perfectly (Figure 2.9C), consistent with their differential contribution to PC2 and PC3 scores. For instance, Nanog occupancy is strongly captured by TF cluster 7 and RNA polymerase binding by TF cluster 11, which have similar contributions to PC3, but opposing contributions to PC2 (Figure 2.8). Overall, the immunofluorescence localization patterns support the spatial segregation of functionally distinct transcriptional and chromatin networks in the ESC nucleus and are consistent with the co-localization of distinct gene regulatory modules detected by our chromosome conformation capture analysis.

2.3.6 - Changes in open and closed chromatin character mirror changes in genome organization during differentiation

Together, our data revealed a close relationship between chromatin character and spatial interactions. The stronger correlation of the 3D chromatin interactome data with PC1, as compared to PC2 and PC3, and stronger description of overall linear chromatin character by PC1

than PC2 and PC3 (Figures 2.6 and 2.8), suggested that the combination of genomic features summarized by PC1 has the strongest association with the organization of spatial chromatin interactions. Based on these results, we predicted that dramatic changes in open/closed chromatin character that occur during differentiation, as defined by positive and negative PC1 scores, respectively, should coincide with strong changes in interactome character.

To test this, we examined interaction differences between ESCs and MEFs with respect to changes in their open/closed chromatin character. To this end, we performed PCA on concatenated ESC and MEF genomes using genomic feature data sets that were available for both cell types (Figure 2.10A; Table 2.6, see *Experimental Procedures*). This generated a new PC1 eigenvector that reflects open/closed chromatin character across both cell types with PC1 scores that are directly comparable between the two cell types (Figure 2.11A-D). Examining the PC1 scores of ten bait regions and their interactomes in both cell types, we found that, as in ESCs, regions with similar PC1 character in MEFs preferentially co-localize (Figure 2.10C). Furthermore, changes in bait character between MEFs and ESCs are generally associated with similar changes in their respective interactomes (Figure 2.10B-D). For instance, the *Dppa2* locus participates in extensive interactions with genomic regions that have positive PC1 scores in ESCs, in accordance with its early replicating and highly transcribed state in pluripotent cells[118]. Upon differentiation, i.e. in MEFs, the *Dppa2* bait region shifts to a PC1-negative, repressed, and late replicating state, and the newly established, MEF-specific interactions likewise have negative PC1 scores, both in *cis* and in *trans* (Figures 2.10B-D, 2.11E). In contrast, the *Rhbdd1* bait region transitions from negative to positive PC1 scores from ESCs to MEFs, and its MEF-specific interactions have significantly higher PC1 scores than the ESC-specific interactions

genome-wide (Figure 2.10C/D). Conversely, bait regions without a change in PC1 character do not change interaction preferences with regards to PC1 scores (Figure 2.10C/D, *Hoxa10* example).

Together, these data support the model where the open/closed chromatin character is the strongest predictor of interaction preferences between distal genomic loci, and that these interactions are subject to change during differentiation in concordance with changes to the PC1 nature of the regions in question (Figure 2.10E). Of note, although the *Pou5f1* gene itself becomes repressed during differentiation[117], the extended 1Mb *Pou5f1* bait region remains strongly positive in PC1 character upon differentiation into MEFs (Figure 2.10C), likely explaining the more limited difference in spatial interactions between ESCs and MEFs for this bait region than for the *Dppa2* locus (Figure 2.5), which shows a more dramatic change in PC1 character upon differentiation.

2.3.7 - The PRC2 component Eed is required for the co-localization of PcG-regulated regions of the genome

Our data described in Figure 2.8 demonstrated that genomic regions are more likely to contact each other when they share strong enrichment of similar regulatory proteins, such as binding by the pluripotency transcription factors Oct4, Sox2, and Nanog with Mediator and Cohesin (represented by negative PC3 scores), or by the repressive Polycomb complex (captured by positive PC3 scores). Therefore, we next considered testing the functional importance of a specific gene regulatory network for long-range chromatin interactions. To this end, we determined chromosome conformation in ESCs lacking the protein Eed, a critical subunit of

Polycomb complex PRC2 that is required for all genomic H3K27me3[128]. We examined particularly the interactions of bait regions with positive PC3 scores, which capture high occupancy by PRC2 and H3K27me3 (Table 2.1). Notably, despite the complete loss of H3K27me3 (Figure 2.12A), *Eed*$^{-/-}$ ESCs continue to express pluripotency-specific transcription factors including Oct4, Sox2, and Nanog at normal levels, maintain their ability to self-renew, and do not spontaneously differentiate when cultured appropriately[10,129], allowing us to test the role of PRC2 without a change in cell identity.

4C-seq analysis showed that the intrachromosomal long-range interactions of the *Hoxd* cluster, a Polycomb-targeted genomic region with highly positive PC3 scores, showed a strong correlation between the *Eed*$^{+/+}$ and the *Eed*$^{-/-}$ ESC lines and extensive long-range interactions in both cases (Figure 2.13A). Despite this similarity, numerous intrachromosomal interactions present in *Eed*$^{+/+}$ ESCs are lost or have a less significant interaction p-value, indicative of a reduced interaction frequency, in the absence of *Eed* (Figure 2.13A, yellow and orange highlights, respectively). We also observed that the regions that show the most dramatic reduction in interaction frequency with the *Hoxd* cluster in *cis* in *Eed* knockout compared to wildtype ESCs often occur at regions of Polycomb enrichment in wildtype ESCs, as defined by high PC3 scores (Figure 2.13A, yellow and orange highlights), and have a PC3 character that is significantly more positive than the chromosome-wide average (Figures 2.13B). These results extend to the interactomes of other Polycomb-regulated regions such as the *Hoxa* and *Hoxb* clusters as well as the *Tbx5* locus (Figures 2.13C-E, 2.12D).

Furthermore, our 4C-seq comparison between *Eed*$^{+/+}$ and the *Eed*$^{-/-}$ ESCs also revealed that the *Hox* clusters interact with each other as well as with other regions of high Polycomb/H3K27me3 enrichment (defined by positive PC3 scores) that encode developmental regulators, in *trans* in wildtype ESCs (Figures 2.13F-H). Importantly, many of these interchromosomal contacts are also lost or reduced in the absence of Eed (Figures 2.13F-H). For instance, the *trans* interactions of the *Hoxb* cluster with the *Hoxa*, *Hoxc*, and H*oxd* clusters are observed in wildtype ESCs, but diminished in knockout ESCs (Figures 2.13F-H). We used FISH to confirm these *trans* interaction data, and consistent with the 4C-seq data, we found that the co-localization frequency between the *Hoxb* cluster and the *Hoxa*, *Hoxc*, and H*oxd* clusters in wildtype ESCs was significantly higher than that in knockout ESCs (Figures 2.13I 2.12B/C).

To explore whether the absence of Eed specifically affects chromatin contacts that occur between those genomic regions that are characterized by positive PC3 scores, we examined the chromatin character of the top ranked long-range interacting regions across five PC3-positive baits (Table 1.1). We found that in *Eed*$^{+/+}$ ESCs, the highest-ranking (and likely most frequent) interactions fall within genomic regions with highly positive PC3 and PC1 scores and with less positive PC2 scores, both in *cis* and *trans* (Figures 2.13J-M, 2.12E). Strikingly, in the absence of *Eed*, the top-ranking interactions no longer occur with regions that are highly positive for PC3 in wildtype ESCs (Figure 2.13J/K), although they still take place between regions with similarly positive PC1 and PC2 scores (Figures 2.13L/M, 2.12E). Thus, the strongest chromatin contacts in the Eed knockout ESCs are no longer between regions with high PC3 status, although they persist between distal genomic regions of similar PC1 and PC2 character, both in *cis* and *trans*. The corollary to this finding is demonstrated by two bait regions that are not enriched for

Polycomb binding (*Pou5f1* and *Ptprg*), as measured by the three principal components: they do not show a dramatic difference in interaction preferences between $Eed^{+/+}$ and $Eed^{-/-}$ ESCs with regards to any of the three principal components (Figure 2.12F-I). These data indicate that loss of PRC2 and H3K27me3 specifically alters the co-association between PC3-positive genomic regions, but does not dramatically affect spatial interactions that relate to the genomic PC1 and PC2 character.

Based these data we conclude that *Eed* is functionally required for the establishment and/or stable maintenance of interactions between Polycomb-occupied, PC3-positive regions in ESCs, both in *cis* and *trans*. Our data also suggest that the overall chromosome topology does not dramatically change upon loss of *Eed* (Figures 2.13A, 2.13L-M). Additionally, the results indicate that regions that were bound by Polycomb proteins and spatially interacting in wildtype ESCs remain spatially confined by a similar chromosome topology in the absence of Eed, but their interaction frequency, i.e. their proximity, is dramatically reduced (Figure 2.13N).

2.4 - Discussion

In summary, our work describes a pluripotency-specific organization of the mouse genome and suggests a model where genes bound by similar regulatory proteins co-localize within the 3D-space of the nucleus. Based on our data we propose a model with two layers of regulation of long-range chromatin contacts in ESCs: (i) We posit that, at the largest scale, the open/closed chromatin character (described by the PC1 character of the genome) defines the regions that have the potential to come into close spatial proximity with one another, both intra- and interchromosomally, which is likely intricately linked to the overall folding of the chromosome

and in agreement with other recent findings[3,124] (Figure 2.6E). (ii) Our data also suggest that, on a finer scale, within the constraints established by the open/closed chromatin nature, genomic regions are more likely to contact each other when they share strong enrichment of similar regulatory proteins, as for example the binding by the pluripotency transcription factors Oct4, Sox2, and Nanog with Mediator and Cohesin (represented by negative PC3 scores), or of the repressive Polycomb complex (captured by positive PC3 scores) (Figure 2.13N). Notably, in addition to our description of PRC2 as critical regulator of specific long-range chromatin interactions (Figure 2.13), a functional requirement for the transcription factor Klf4 in the maintenance of long-range chromatin contacts in ESCs has recently been reported[102], extending previous findings that demonstrated a requirement for Oct4 in the organization of short range chromatin interactions within the extended *Nanog* locus[95].

Our results also demonstrate that the depletion of a single gene regulatory network in ESCs specifically affects long-range interactions of genomic regions particularly strongly enriched for occupancy by this factor (in our case PRC2 and H3K27me3), without altering the global interaction network generated by the open/closed chromatin character. A potential explanation for the limited effect on overall chromatin interactions may be that the chromosomal conformation chassis is likely maintained by many combinatorially acting regulatory factors that probably involve numerous interactions mediated by Cohesin[119,130].

The role of cell-type specific gene regulatory networks in defining specific long-range chromatin interactions potentially allows TADs enriched for specific gene regulatory features to co-localize in the 3D space of the nucleus. Notably, while our results demonstrate that Polycomb

complexes are important for long-range chromatin contacts associated with the regulation of Polycomb targets in mammals, it has recently been shown that the TAD structure within a specific locus, the X-chromosome inactivation center, is not affected by the *Eed* knockout[43], indicating different regulatory mechanisms at the different hierarchies of genome organization. It is also interesting to speculate that the interaction between the *Hoxd* and *Hoxc* clusters may provide a mechanism for how the non-coding RNA Hotair, encoded within the *Hoxc* cluster, finds its target genes within the *Hoxd* cluster located on a different chromosome[131], by exploiting 3D conformation of the genome, in a similar manner to that employed by long-noncoding RNA *Xist*[132].

Based on our data and other reports[50,105], we suggest that cell type-specific transcriptional networks participate in preferential spatial interactions within a general chromosome folding interaction network in all cell types. The preferential co-localization of distal genomic regions with similar regulatory networks tens of mega-bases apart on the same chromosome, and in *trans*, suggests a previously unappreciated role for transcriptional networks in influencing the 3D positional preferences of chromatin in mammalian cells, which may represent an evolutionarily conserved interaction between eukaryotic genome organization and gene regulation[4,133]. We speculate that this organizational hierarchy facilitates the recruitment of regulatory proteins to their genomic target sites and the establishment of local chromatin environments, which are both critical for the efficient regulation of gene expression.

CHAPTER 3

Discussion, outstanding questions, and future directions

3.1 - Open and closed chromatin as the backbone of genome organization

We (Chapters 2.3.3 and 2.3.6) and others[2,3,36] have demonstrated that long-range chromatin contacts occur predominantly between distinct parts of the genome defined by their open and closed chromatin states. Active, accessible chromatin associates with active, accessible chromatin, and closed, inaccessible chromatin with closed, inaccessible chromatin. Why is this the case?

Association of transcriptionally repressed DNA to the nuclear lamina (reviewed in Chapter 1.3) is a strong contender for the molecular force partitioning of the genome into an active and inactive compartment. Nature has provided numerous tests of this system in both development and disease. H3K9me3-enriched chromatin is targeted to the nuclear periphery via an interaction between the chromoshadow domain of heterchromatin protein 1 (HP1) and the nuclear envelope-associated Lamin B Receptor (LBR)[134]. Interestingly, LBR expression is developmentally regulated in mouse retinal cells, where it is lost by post-embryonic day 14, and replaced by LaminA, which maintains the peripheral localization of heterochromatin. Remarkably, in Rod cells, LaminA expression is not initiated, and loss of LBR expression results in the re-localization of condensed chromatin from the nuclear periphery to the nuclear center[135]. This demonstrates that tethering of condensed, closed chromatin to the nuclear periphery is not required for the compartmentalization of open and closed chromatin.

Alternatively, co-localization may be the default state of silenced chromatin due to the self-associating nature of individual heterochromatin components[136,137]. In favor of this model, insertion of heterochromatic satellite repeats into *Drosophila* euchromatin induces co-

localization of the targeted locus and endogenous heterochromatin[138]. This model also predicts that long-range interactions should be promiscuous in the absence of condensed, closed chromatin, as genome compartmentalization would be lost. Intriguingly, loss of AtMORC1 and AtMORC6 in *Arabadopsis*, proteins required for the silencing and compaction of methylated DNA, but not DNA methylation itself, results in the loss of genome compartmentalization[139].

Potentially at odds with this model, ESCs, which lack extensively co-localized compacted chromatin[55], still demonstrate detectable compartmentalization of the nucleus that correlates with features enriched in euchromatin[42] (Chapter 2.3.3). I speculate that numerous chromatin remodelers, including Chd1 and Brg, which are highly expressed in ESCs and act in concert to maintain pluripotency[54,140], compete with the spread of heterochromatin in ESCs, resulting in both the transcriptionally permissive milieu of the ESC nucleus, and the closed-chromatin based compartmentalization of the ESC nucleus. This balance may contribute to the ESC-specific organization of the mouse genome (Chapter 2.3.2), as well as to the developmental plasticity ESCs. Re-localization and/or down-regulation of these chromatin remodelers may result in the condensation of specific regions of lineage-specific chromatin and the establishment of a lineage specific chromosome conformation. Working in tandem, numerous Cohesin[119,130]-mediated long-range chromatin interactions among distal transcription factor bound loci may act in concert to spatially confine open chromatin.

3.2 - Long-range chromatin interactions and transcription

The identification and analysis of co-localized genomic regions among the open and accessible compartment of the genome has demonstrated the existence of an interaction landscape existing

on top of the genome compartmentalization discussed in Chapter 3.1. The study of this interaction landscape across various cell types (Chapter 2.3.4, Chapter 1, and below) has allowed for the identification of *trans* acting factors that may have causal roles in genome organization.

The role specific transcriptional networks play in genome organization, above and beyond the preferential co-localization of expressed genes, has been a challenging area of study, due to the difficulty of dissecting out the overlapping contributions of genome compartmentalization and gene regulatory components in organizing chromatin. For instance, in fetal liver cells, the Klf1 regulated beta-globin gene locus preferentially interacts with genomic regions enriched for Klf1 DNA binding motifs, and with genes up-regulated by Klf1 at transcription factories enriched for Klf1. In this system *Klf1* deletion results in the loss of co-localization of candidate Klf1-target genes by DNA fluorescence in situ hybridization (FISH), however the expression of target genes is also reduced by the *Klf4* deletion[51], making it difficult to assess if the candidate loci remained associated with the same chromatin compartments. Insertion of a transgenic human beta-globin locus control region (LCR, a series of enhancers controlling the regulated expression of the globin locus[141]) into the mouse genome demonstrated that the LCR does not effect the interactome of the targeted locus. However, contacts between genomic regions enriched for GATA1- and EKLF-regulated genes (core transcription factors regulating erythropoiesis[142]) were strengthened, arguing for the role of this cell type-specific transcription network in the organization of intra-compartmental interactions. Notably, the transgenic LCR was only able to up-regulate expression of genes located in *cis* to the endogenous LCR, and not other contacted genes[50], demonstrating importantly that co-localization of genes bound by the same regulatory proteins is insufficient for a distal DNA

regulatory element to exert influence over gene expression. Similar findings in other systems have also been described, such as: the T-cell lineage, where the T-helper cell (Th1) *Ifng* interactome, upon differentiation from naïve T-cells, strengthens a sub-set of pre-existing chromatin contacts that contain Th1-specifically expressed genes and Stat4 binding sites[106]; the strengthening of pre-established chromatin interactions between glucocorticoid receptor bound DNase hypersensitive sites upon receptor stimulation[105]; as well as E2A dependent interactions between E2A bound loci within the open compartment of in pro-B cells[143].

Although these findings do suggest enrichments for specific transcriptional programs within chromatin interaction networks, it is difficult to untangle the contribution of specific transcriptional networks in guiding chromatin interactions from the general preference for transcribed, open regions of the genome to co-localize in nuclear space. More recent work has shed additional light on the ESC-specific organization of the mouse genome and the genomic features upon which it relies. In line with their requirement for establishing sub-megabase-scale chromatin contacts, Cohesin and Mediator are required for the maintenance of ESC-specific chromatin contacts between the *Nanog* locus and the remainder of the genome[119]. Additionally, Nanog and Oct4 (both of which bind DNA with Mediator and Cohesin[91]) are required for interactions between genomic regions normally highly enriched for these transcription factors. Furthermore, ectopic recruitment of Nanog protein to chromatin can induce chromatin contacts in *cis* with other Nanog enriched genomic regions, within the context of a pre-existing chromatin conformation[144]. Effects on the transcriptional status of genes within the ectopically induced loops were not discussed in that work.

Cell type-specific, long-range chromatin contacts within the closed chromatin compartment have also been reported. Olfactory receptor (OR) gene expression is monogenic in individual olfactory sensory neurons, such that each neuron expresses only one OR, transcribed from a single allele[145]. Monogenic expression is achieved, in part, through the cell type-specific and LBR-dependent aggregation of non-transcribed[146], H3K9me3/H4K20me3 silenced[147] OR genes into a small number of distinct heterochromatic foci[146]. These foci contain OR genes from multiple chromosomes, and exclude the actively transcribed OR allele[146]. Notably, silenced OR genes in other cell types within the optical epithelia do not demonstrate OR-specific clustering, and instead show a dispersed localization of OR genes throughout nuclear heterochromatin, suggestive of an olfactory receptor neuron-specific mechanism for the clustering of OR genes within the closed chromatin compartment[146].

These findings, along with our work on the influence of ESC transcriptional networks on genome organization (Chapters 2.3.4, 2.3.5, and 2.3.7), have made it clear that chromatin contacts preferentially occur between regions of shared gene regulatory-factor binding within a pre-defined open chromatin space in the nucleus.

Taken together, this paradigm converges on a profoundly simply yet difficult question in the field of mammalian genome organization: Are long-range chromatin interactions involved in gene regulatory processes?

Regulatory interactions between enhancers and promoters in *cis* can occur over long distances. For instance, the limb-specific *Sonic hedge hog* (*Shh*) enhancer ZRS, whose mutation

results in polydactyly[148], resides over 1Mb away from *Shh* and regulates its expression via chromatin looping[149]. Scores of long-distance enhancer-promoter interactions have been described[150], however few of those characterized, if any, occur over distances larger than the largest TAD (~5Mb, median size ~880kb[42]), suggesting that enhancer action, and by extension activating, gene-regulatory chromatin interactions, may be limited to intra-TAD space[45].

Transvection, the pairing of homologous chromosomes in interphase, is a common phenomenon in *Drosophila* allowing DNA elements to exert gene regulatory effects in *trans*[151]. Likewise, in *Drosophila*, DNA encoded Polycomb Response Elements function in *trans* in the absence of strong homology[152] (discussed in detail in Chapter 3.3). Similar *trans* acting chromatin interactions have proven to be elusive in mammalian systems.

Long intergenic non-coding RNAs (lincRNAs) potentially provide an example of *trans* acting, chromatin-associated gene-regulators. In an expansive study, knockdown of 137 out of 147 lincRNAs identified in ESCs resulted in significant gene expression changes among protein coding genes[153]. Remarkably, the majority of the effected genes are coded for at distances greater than 300kb from the site of lincRNA transcription[153], arguing that lincRNAs function over exceptionally long distances. *HOTAIR* is perhaps the best characterized *trans* acting lincRNA. Transcribed from the *Hoxc* locus (chromosome 15), it interacts with the PRC2 complex and is required for H3K27me3 of the *Hoxd* cluster (chromosome 2)[131]. Although this lincRNA functions in *trans*, our work (Chapter 2.3.7) suggests that this *trans* action may be mediated by the juxtaposition of the *Hoxc* and *Hoxd* loci, adding to emerging evidence that non-coding RNAs act over long distances through chromatin contacts[132].

Testing the requirement for long-range *cis* and *trans* interactions in the regulation of transcription will prove to be extremely challenging, as every factor identified to date that has an influence on, or is associated with, long-range chromatin contacts has a central role in mediating the effects of *cis* regulatory elements, likely ruling out loss of function experiments. Gain of function experiments via insertion of transgenic regulatory elements or recruitment of regulatory proteins to chromatin via insertion of Lac operators and Lac repressor-fusion proteins, similar to approaches taken by the de Laat group[50,144], will be useful gain of function studies. Alternatively, regions found to co-localize based on transcription network membership (Chapter 2.3.4) could be tested with a DNA/RNA FISH approach, where frequency of transcription (presence of an RNA signal) could be compared directly to interaction frequency (DNA FISH signal) for a set of genomic loci, enabling one to ask if transcription occurs more frequently when two loci are interacting than when they are not.

3.3 - Long-range chromatin interactions and transcriptional repression - Polycombs

In *Drosophila*, PcG proteins are recruited to Polycomb Response Elements (PREs)[154]. Although putative PREs have been identified in mammals[155,156], it is unclear if PcG-proteins are recruited to mammalian chromatin via PREs as a rule. *Drosophila* PREs regulate the repression of target genes over large genomic distances in *cis* as well as in *trans,* and interestingly, have little repressive effects on their own, but instead require multiple PREs acting in concert[152]. For instance, transgenic copies of *fab-7* PRE co-localize, both in *cis* and in *trans,* and repress transcription of genes proximal to PRE insertion sites[157]. Unlike transvection-based *trans* gene regulatory processes in *Drosophila*, PRE co-localization does not require a high degree of DNA

sequence homology[152]. PREs exert their effects, in part, through tissue-specific Polycomb-[126] and RNAi machinery-[158] dependent co-localization in nuclear space. PRE co-localization nucleates the formation of Polycomb bodies, dynamic sub-nuclear structures enriched for PcG components[126,159] which are conserved in mammals (Chapter 2.6). Demonstrating robust function in *trans*, deletion of the endogenous *fab-7* PRE leads to de-repression of genes proximal to *fab-7* transgenes. This de-repression is inherited through cell division, arguing for a role for PRE co-localization in the maintenance of epigenetic memory[157].

Recent work utilizing Hi-C in fly embryos has demonstrated that genomic regions enriched with PcG-protein binding co-localize in general, extending this gene regulatory paradigm beyond well-characterized PREs[4]. In *Drosophila,* PcG-protein dependent co-localizations are limited in scope within a fixed underlying chromosome conformation[127], similar to the genome structure describe in Chapter 2 for mammals, where specific transcriptional networks fine-tune chromatin contacts within the context of an open/closed chromatin-based chromosome conformation. Interestingly, Polycomb-dependent chromatin contacts between the two *Drosophila Hox* clusters, whose mammalian homologues are studied in detail in Chapter 2.3.7, have also been reported, and are conserved between *D. melanogaster* and *D. virilis*[52].

The role of PcG proteins in mammalian genome organization is less clear. The developmentally regulated *Gata4* and all four *Hox* loci are organized into complex looping structures spanning a few hundred kilobases in *cis*. These looped structures are disrupted to a small degree upon knockdown of PRC2 component Ezh2, but appear to rely more strongly on the transcriptional state of the respective loci[160,161]. The open chromatin compartment of mouse

pro-B cell nuclei is segregated to some degree by a depletion of interactions between chromatin enriched for H3K4me2 and chromatin enriched for H3K27me3[143], however this study did not differentiate between intra-TAD and inter-TAD interactions making this result difficult to interpret due to the tendency of H3K27me3 to be found in large domains along linear chromatin[162].

In Chapter 2.3.7 I showed that Polycomb repressed regions of the mouse genome co-localize in 3D space, that these preferential co-localizations occur within the context, but do not influence to structure of, an existing chromosome conformation, and that these interactions rely to some extent on the PRC2 complex. Neither the self-renewal of pluripotent cells or the expression of pluripotency transcription factors is compromised upon loss of PRC2[10,129].

Moving forward, genetic manipulation of either whole *Hox* clusters, or the putative mammalian PREs will be informative with regards to the requirement for *trans* interactions in the repression of Polycomb targets.

3.4 - Concluding remarks

In this work we identified a pluripotency-specific organization to the mouse genome, showed that genomic regions enriched for binding of distinct transcriptional network members co-localize with similarly bound regions (including OSN and Polycomb networks in ESCs), visually confirmed the spatial segregation of the mouse genome into Nanog- and Polycomb- enriched regions using immunofluorescence, and demonstrated a requirement for Eed in the preferential interactions of Polycomb regulated genomic regions.

Although the spatial segregation of open chromatin into OSN bound and Polycomb bound is clear (Chapter 2.3.4), as is the preferential co-localization of loci regulated by transcriptional networks in other cell types (Chapter 3.2), the requirements for and role of this segregation in gene regulatory processes are not. Future work using the approaches touched upon above should help to answer the questions posed in this chapter, and help us to more fully understand the role of genome organization in gene regulation.

CHAPTER 4

Experimental Procedures and Analysis

4.1 - 4C Library Preparation

4C libraries were prepared from mouse ESCs (V6.5 line), MEFs (wildtype of 129SvJae background), an iPSC line (described in Chin, Plath et al., manuscript in preparation), the pre-iPSCs (12-1 line)[163], and the *Eed* mutant ESC line (17Rn5-3354SB) and a sibling wild-type ESC line[164], essentially as described[1]. Specifically, 10^7 cells were trypsinized and filtered to single cell suspensions with 40um cell strainers. Following a PBS wash, cells were cross-linked in 1xPBS with 10% fetal bovine serum (FBS) and 2% formaldehyde. Ice-cold 1M glycine was added to the cells on ice to a final concentration of 0.13M to quench the crosslinking reaction. Cells were spun at 500g for 5 minutes at 4°C, resuspended in ice-cold lysis wash buffer (10mM Tris-HCl pH7.5, 10mM NaCl, 5mM $MgCl_2$, 0.1mM EGTA, and 1x protease inhibitors (Roche)), and re-pelleted at 4°C. Pellets were subsequently resuspended in 1ml ice-cold lysis buffer (lysis wash buffer with 0.2% IGEPAL (Sigma)), cells were lysed on ice for 30 minutes, and dounced using a tight piston for 10 strokes to isolate nuclei. Nuclei were spun down at 200g for 7 minutes at 4°C, washed with 1.2x buffer B (Roche), resuspended in 1ml ice-cold 1.2x buffer B, and transferred to non-stick tubes. Cells were brought to 37°C and 20ul 15% SDS were added to each tube, which were then incubated for 1 hour at 37°C while rotating end over end. 150ul Triton X-100 was added and tubes were allowed to incubate for another hour at 37°C. 800U of high concentration HindIII (Roche, cat# 10798983001) were added to each tube and the restriction digest reaction was run overnight at 37°C while tubes rotated. The restriction enzyme was inactivated at 65°C for 25 minutes and digest efficiency was determined as described[1]. Digested samples were transferred to 50ml falcon tubes, and 5.3ml H_2O and 700ul 10x ligation buffer (660mM Tris-HCl pH 7.5, 50mM $MgCl_2$, 50mM DTT, 10mM ATP) were added to each sample. 100U T4 ligase (Roche cat#10799009001) were added and samples ligated overnight at 16°C.

Ligation efficiency was checked as described[1]. 30ul 10mg/ml proteinase K were added to efficiently ligated samples and the samples incubated overnight at 65°C. Subsequently, 30ul RNaseA (Invitrogen, cat# 12091-021) were added to each sample, and samples were incubated for a additional 45 minutes at 37°C. DNA was phenol-chloroform extracted, and precipitated by the addition of 7ml H_2O, 1ml 3M Na-acetate pH 5.6, 7ul glycogen (Roche, cat# 10901393001), and 35ml 100% ethanol, followed by freezing at -80°C. Precipitated DNA was spun down at 8800g for 45 minutes at 4°C, washed with ice-cold 70% ethanol, and re-spun at 3000g for 15 minutes at 4°C. Upon drying, DNA was resuspended in 150ul 10mM Tris pH 7.5, 300ul H_2O, 50ul 10x DpnII restriction buffer (NEB), and 50U high concentration DpnII (NEB, cat# R0543M) were added to each tube and the DNA was digested again overnight at 37°C. DNA was phenol-chloroform purified, precipitated via addition of 50ul 3M Na-Acetate and 1ml ethanol, and re-dissolved in 100ul H_2O. The DpnII-digested DNA was then transferred to a falcon tube to which 12.5ml H_2O, 1.4ml 10x ligation buffer, and 200U high concentration T4 ligase were added for ligation overnight at 16°C. Following phenol-chloroform purification, samples were precipitated via the addition of 14ul glycogen and 35ml 100% ethanol at -80°C. The precipitated DNA was pelleted and washed as above, and resuspended in 150ul 10mM Tris-HCl pH 7.5. Residual salt was removed via Qiagen PCR purification columns.

4.2 - 4C library PCR amplification and Illumina high-throughput sequencing

Inverse PCR primers (Table 2.2) were designed to anneal to a bait locus HindIII/DpnII restriction fragment (selected with the criteria that it be longer than 300bp and be within 50kb of the indicated gene) and to amplify the unknown portion of the chimeric DNA circle generated during 4C library preparation. The resulting DNA circles consist of the bait locus restriction

fragment and its interacting partner's restriction fragments. The six 3' nucleotides of the primers annealing to the HindIII side of the restriction fragment contained the HindIII restriction site when possible, or were generally within 4bp of the start of the HindIII site, to avoid uninformative reads upon sequencing. Primers on the DpnII side were allowed more positional flexibility, as sequencing data were not produced from the DpnII end of the restriction fragment.

200ng of DNA from the 4C library were used as template for the PCR amplification using the Expand Long Range PCR system (Roche). 5uM each of forward and reverse primer lacking Solexa sequencing adaptors were applied to amplify the interactome of interest in a 25ul reaction volume under the following PCR conditions: 1 cycle at 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; 1 cycle of 68°C 7 minutes. PCR products were run on an agarose gel, and amplicons between 100-500bp were isolated, gel extracted (Qiagen Gel purification system), and used as template for a second PCR reaction utilizing the same primers with the addition of the Solexa adaptors in a 50ul volume as follows: 1 cycle of 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; (92°C 30 seconds; 68°C 1 minute +20 seconds/additional cycle; 68°C 1 minute) x 15 cycles; and 1 cycle at 68°C for 7 minutes. The PCR-amplified library was purified over GFX PCR DNA purification kit columns (GE Healthcare) to remove primer dimers, followed by a second purification with Qiagen MinElute Reaction Cleanup kit (Qiagen) to remove residual salt. Samples were quantified using the Quant-iT dsDNA BR assay kit quantification system (Invitrogen) with a Qubit fluorometer. Purified, PCR-amplified 4C-seq libraries were pooled in EB (Qiagen), 0.1% Tween-20 for multiplexed sequencing as primer distinctiveness allowed, and sequenced at the Broad Stem Cell Research Center at UCLA.

4.3 - 4C sequencing and read mapping

Two to seven 4C libraries were multiplexed and sequenced using the Illumina Genome Analyzer II to obtain 76 base pair (bp) reads or Illumina HI-seq-2000 to obtain 100 bp reads. Reads were parsed based on a unique, non-annealing two base pair bar code and/or unique bait-specific primer sequences. The resultant reads were mapped to the mouse genome (build MM9) using Bowtie software[165]. Only reads that aligned to a unique position in the genome with no more than two sequence mismatches were retained for further analysis. Reads that were successfully aligned to the genome were then remapped to the 736,199 unique HindIII sites along the genome by matching their respective loci. Because we were concerned about any potential biases created by differential mappability, we excluded all HindIII sites that do not precede a unique 50bp along both DNA strands. In other words, we only mapped reads to HindIII sites that are unique in the MM9 genome with respect to the 100bp centered on the hexamers that comprise the sites, as illustrated below:

(+) AAGCTT--- 50bp ---

(-) --- 50bp ---TTCGAA

By only considering unique HindIII sites, we have restricted our 4C analysis to highly mappable regions of the genome

Table 2.1 provides a summary of all the bait regions for which 4C-seq libraries were generated for all cell types discussed in this study. Read distributions and statistics for all individual data sets that passed quality control steps (see below) are contained in Table 2.3.

4.4 - 4C hit determination

To reduce potential clonal amplification effects inherent to PCR-based genomic approaches, we collapsed the raw 4C-seq read count at each HindIII site down to a 'hit' if the site met a read count threshold. The threshold was chosen so that at least 80% of all hits were intrachromosomal, i.e. at most 20% of our hits fell in *trans* (Figure 2.1E). For each library, the threshold and the number of sites that passed this threshold (i.e. the number of hits) are provided in Table 2.3. For the majority of libraries the count threshold for calling hits was a single read.

4.5 - 4C library quality control (QC)

We used three criteria adapted from van de Werken et al.[166] to estimate the quality of our individual 4C-seq libraries, as well as an additional two criteria of our own. First, we checked whether the library under consideration was comprised of at least 500,000 reads in total. Second, the *cis*/genome read count ratio (the number of mapped reads in *cis* over the total number of mapped reads) had to be at least 20%. Third, at least 20% of HindIIII sites within the 2Mb region around the bait had to be covered by at least one read. If a library passed all three of these de Laat group-inspired criteria, then it received one credit. Libraries received additional credit for passing each of the following two tests: 1) having a *cis:trans* hit ratio that was at least 4:1 (i.e. a maximum of 20% of thresholded hits could fall in *trans* for the library to be credited); and 2) having at least 1.5% of all sites along the *cis* chromosome covered by a hit. In summary, each library could achieve a maximum score of 3/3. Libraries that received a total score of 3/3 automatically passed QC, while those with 2/3 were subject to further scrutiny and only passed if they exhibited strong metrics for their two passing criteria. The libraries scoring less than 2 did not pass QC. Excluding three control libraries, 198/242 (84%) of the 4C-seq libraries passed the

QC and were kept for further analysis (Figure 2.1D). Those libraries that passed QC are given in Table 2.3.

To ensure that those genomic regions within closed chromatin environments (negative PC1 scoring regions) were digested as efficiently as those regions in open chromatin surroundings, we required two things: 1) Our 4C-seq libraries had to exhibit at least 20% HindIII site coverage within the 2Mb regions around the bait, as shown in Table 2.3 and Figure 2.1D. Notably for baits with negative PC1 scores, this demonstrates that closed chromatin is subject to proper digestion. 2) Regions proximal to the bait of high and low PC1 character had to show similar average hit probability (data not shown).

4.6 - Pooling of replicate 4C-seq libraries

Replicate 4C-seq libraries for a single bait locus in a given cell type that passed QC (described above) were pooled by calculating the average number of times each site was called a hit (by the thresholding criteria described above) in all replicates. In essence, we determined the probability of a hit at each HindIII site along the genome for each bait and cell type. Table 2.1 summarizes the bait regions for which pooled 4C-seq libraries were produced in each cell type, while Table 2.3 lists all the replicate libraries that were pooled, as well as statistics pertaining to hit probability. In this study we considered 66 pooled 4C-seq data sets (Table 1.1).

4.7 - Definition and calculation of 4C hit percentage

In order to obtain a smoother continuous signal at a scale that was compatible with our genomic feature and PCA data (see below), we determined the average hit probability within 200 kilobase

(kb) windows tiled along each chromosome, referring to this as the 'hit percentage.' We observed a strong correlation between hit percentage and binned Hi-C read counts using equivalent window sizes (Table 2.5).

4.8 - Binomial test analysis

To demarcate positively interacting regions of a bait locus along each chromosome, we sought to identify statistically significant clusters of HindIII sites that exhibited a high probability of being hit across replicates. We used R's binomial test function[167] to calculate the probability of seeing the observed proportion of hits to HindIII sites, or 'hit percentage', within a 200kb window around each HindIII site along each chromosome, relative to the expected proportion obtained by modeling the average hit percentage as a function of distance from the bait locus across all data sets in a given cell line (see below for details on the background modeling) in *cis*, or relative to the average hit percentage for each respective *trans* chromosome. Using the observed hit probability within the 200Kb surrounding each site, we determine the number of hits that this represented, given the number of sites within the window: i.e. number of hits = hit probability * number of sites with 200Kb window. We used the resulting number of hits as the value for the binomial test parameter representing the number of successes, with the number of HindIII sites being the number of trials, and the hit percentage obtained from the empirical background model (for a locus at the given distance from the bait locus in *cis*) or the average hit percentage in *trans* being the hypothesized probability of success. Only sites centered within windows containing at least ten HindIII sites were considered. A threshold of $-\log_{10}$(p-value) >= 1.8 was used to determine HindIII sites centered within windows showing significant clusters of interaction. This threshold was used as it resulted in a small false discovery rate (FDR; see determination below)

in *cis*, while allowing us to pick up significant *trans* interactions. The binomial test results are given in Table 2.4. Adjacent and overlapping positive windows were concatenated into 4C-positive domains. Table 2.1 catalogs these significantly-interacting domains for pooled libraries ("interactome") as determined by our 4C-seq analysis pipeline.

In order to determine an FDR for each of our 4C-seq interactomes, we generated corresponding data sets of simulated hit probabilities. For each bait, we generated as many data sets of simulated hits as we had experimental replicates for that bait as follows: To simulate intrachromosomal hits we used the hit percentage specified by the empirical background model (described below) as the probability of sampling a hit at each site with respect to its distance from the bait. To simulate interchromosomal hits, the average hit probability for each chromosome in the experimental pooled data set was used as the probability of sampling a hit at each site along the chromosome. We pooled the resulting simulated replicates in the manner described above for our experimental replicates. Table S4 lists the number of significant windows and FDR in *cis* and *trans* for each of our 4C-seq data sets.

A 200kb window-size for the binomial test was chosen after having tested various window sizes and generated a multi-scale representation of the results (domainogram) to confirm the consistency of the binomial test p-values across window sizes (data not shown). This window size contained a sufficient number of HindIII sites (trials) to produce a robust binomial test result, while providing sub-megabase resolution.

To confirm that there was no strong bias for higher mappability within our libraries, illustrating that our selection of unique HindIII sites (as described above) had the desired effect of only interrogating highly mappable regiions of the genome, we compared the average mappability within our 4C positive regions (4C domains) to that within 4C- negative regions, for intra-chromosomal interactions. We downloaded the UCSC ENCODE MM9 mappability scores based on 36bp alignment (ftp://encodeftp.cse.ucsc.edu/pipeline/mm9/wgEncodeMapability/wgEncodeCrgMapabilityAlign36mer.bigWig), using the 36mer mappability data because it reflects the typical length of our 4C-seq reads after trimming of the primer prefixes. The 36mer-based mappability scores are also the most stringent of the available mappability scores. All intra-chromosomal interactomes produced in our study are shown in the bar plot of the average difference in mappability scores (mean score of 4C domains (inside) - mean score of outside 4C domains (outside)) for all 66 of our baits (Figure 2.1G). We found that the average difference in mappability score for all baits (labeled 'MEAN') is small, supporting the conclusion that our results are not significantly biased by mappability. A similar lack of mappability bias was observed when genome-wide interactions were examined (data not shown). This result is consistent with the fact that we only considered data at unique HindIII sites in the genome.

4.9 - Intrachromosomal empirical background model

To obtain an estimate of the expected probability of success for the binomial test, we used the average hit probability within 200kb windows from the bait locus across all baits to build a regression model of hit probability as a function of distance to the bait for each cell type. Adapting the approach used by Lieberman-Aiden et al.[3] for their Hi-C data, we used the 1 – 8

Mb region proximal to the bait locus to fit a log-log regression model. The data in this region produced a much better fit than using the entire range of data points available (Table 2.7). We also noticed that within 1 Mb of the bait, the data deviated sharply from the power law scaling observed over the 1 – 8 Mb region, and that a linear regression model was more appropriate for this immediately proximal region around the bait locus. We therefore performed segmented regression analysis by partitioning the distance from the bait locus into the 0 – 1 Mb region and the 1 – 8 Mb region, using linear and log-log regression on these two regions, respectively. Model parameters and R-squared values are provided in Table 2.7. Figure 2.3A,iii gives an example of the resulting empirical background model generated for the ESC *Pou5f1* pooled data set.

4.10 - Data set correlation, overlap determination, and clustering

To compare interactomes across cell types, we performed Spearman's rho correlations on hit probabilities within 200kb windows tiled across each chromosome, and calculated Jaccard similarity coefficients using binary vectors representing 200kb tiled windows that overlapped 4C-positive domains (Figures 2.5 and 2.2B). Dendrograms were obtained by converting the Spearman's *rho* statistics and Jaccard coefficients into distance measures and performing unsupervised hierarchical clustering using R's hclust function[167] using the Ward agglomeration method.

4.11 - RNA-seq

Strand-specific RNA-seq from V6.5 ECSs and wildtype 129SvJae MEFs (Table 2.6) was performed essentially as described in Parkhomchuk et al[168], using 4 ug of total RNA as starting

material. Reads were mapped to the mouse genome (MM9) using TopHat software[169] and only those reads that aligned with no more than two sequence mismatches were retained.

4.12 - ChIP-seq

All histone modification data used for this study (Table 2.6) were determined using native ChIP[170]. Briefly, nuclei were isolated from non-crosslinked V6.5 ESCs and 129SvJae MEFs by centrifugation through a sucrose cushion (1.2M sucrose, 60mM KCL,15mM NaCl, 5mM $MgCl_2$, 0.1mM Tris-HCl, 0.5mM DTT and protease inhibitor cocktail). Nuclei were then resuspended in Mnase-digestion buffer (0.32M sucrose, 50mM Tris-HCl, 4mM $MgCl_2$, 1mM $CaCl_2$, and protease inhibitor cocktail) and digested with 3 units of MNase (Roche) for 10 minutes at 37ºC. Soluble chromatin fractions were incubated with anti-H3K4me3 (Abcam; ab8580), anti-H3K4me2 (Abcam ab7766), anti-H3K4me1 (Abcam; ab8895), anti-H3K27me3 (Active Motif; 39155), anti-H3K27ac (Abcam; ab4729), and anti-H3K36me3 (Abcam; ab9050), respectively. Extracts were washed twice with wash buffer A (50mM Tris-HCl, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl, 10mM EDTA, 125mM NaCl), wash buffer C (50mM Tris-HCl, 10mM EDTA, 250mM NaCl), once with LiCl buffer, and once with 1xTE. DNA extraction and library preparation as described[170].

Transcription factor binding data generated in this study (Table 2.6) were acquired using cross-linking ChIP. V6.5 ESCs and 129SvJae MEFs were grown to a final concentration of $5x10^7$ cells for each sequencing experiment. Cells were chemically cross-linked by the addition of formaldehyde to 1% final concentration for 10 minutes and quenched with glycine at a final concentration of 0.125 M. Cells were then resuspended in buffer I (0.3M sucrose, 60mM KCl,

15mM NaCl, 5mM $MgCl_2$ 10mM EGTA, 15mM Tris-HCl, 0.5mM DTT, 0.2% NP-40, and protease inhibitor cocktail), and incubated on ice for 10 minutes. Nuclei were generated by centrifugation in a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM $MgCl_2$, 0.1mM Tris-HCl, 0.5mM DTT, and protease inhibitor cocktail). Isolated nuclei were resuspended in sonication buffer (50mM Hepes, 140mM NaCl, 1mM EDTA, 1% TritonX-100, 0.1% Na-deoxycholate, 0.1% SDS), and sonicated using a Diagenode Bioruptor. Subsequently, nuclear extracts were incubated overnight at 4C with one of the following antibodies: anti-Klf4 (R&D; AF3158), anti-Myc (R&D; AF3696), anti-Nanog (Cosmobio), anti-Oct4 (R&D; AF1759), anti-Sox2 (R&D AF2018), anti-p300 (SantaCruz; sc-585). Extracts were washed twice with RIPA, low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1xTE. Reverse cross-linking was performed by overnight incubation at 65C with 1% SDS and proteinase K. All protocols for Illumina/Solexa sequencing library preparation, sequencing, and quality control were performed as recommended by Illumina, with the minor modification of limiting the PCR amplification step to 10 cycles.

Reads were mapped to the mouse genome (MM9) using Bowtie software[165] and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the same location in the genome were collapsed to a single read to account for clonal amplification effects. ChIP-seq peaks were called using MACS software (Version 1.4.2)[171] using a bandwidth parameter of 150 bp.

4.13 - Chromatin states

Five chromatin states in ESCs and MEFs were identified at a resolution of 200bp as described by Ernst and Kellis[123] using the six histone modification ChIP-seq data sets listed in Table 2.6, plus a ChIP input dataset.

4.14 - Transcription factor clusters

The genome was tiled into 1kb windows and the presence of transcription factor (TF) peaks from sixteen in-house and previously published ChIP-seq data sets for ESCs were used to define the TF clusters used in this analysis (Figure 2.6A and Table 2.6)**.** For published data sets, we used peaks determined by the authors of the respective studies. The "Cohesin" data set represents the merging of peaks from the Smc1 and Smc3 data sets, and the "PRC2" data set represents the merging of peaks from the Eed, Ezh2, and Suz12 datasets (Table 2.6). This procedure resulted in a vector of binary data for each TF reflecting its absence or presence within 1kb windows across the genome. The windows represented by these vectors were then clustered using R's k-means function using the Hartigan-Wong method[167] to obtain groups of windows exhibiting common combinatorial binding patterns across the genome. The number of centers (k=11) was chosen so as to substantially reduce the number of potential combinatorial TF groups (2^{16}-1), while ensuring that each cluster was represented by a significant number of windows (>7000 windows or ~0.25% of the genome). The eleven combinations of TFs found to co-bind within each window of a cluster are analogous in their combinatorial nature to the five chromatin states described above.

4.15 - Principal Component Analysis

To compare our 4C interactome data to linear genomic features, including gene density, gene expression, replication timing, chromatin states, and transcription factor combinations in ESCs, we used Principal Component Analysis (PCA) to reduce the dimensionality of the 29 linear ESC genomic feature data sets (Table 2.6; discussed above and below). This allowed us to focus on weighted combinations of features, or principal components (PCs), that best characterized the genomic landscape of a cell type. Chromatin states and transcription factor clusters were used in order to capture the biologically important combinatorial nature of these features. The five described 200bp-based chromatin states and eleven 1 kb-resolution transcription factor cluster data were binned into 200kb windows resulting in a semi-quantitative profile of feature density across the genome. Density profiles for DNase and LaminB were obtained in a similar manner by tallying the number of times they were present (at 1kb resolution) within 200kb windows. Replication timing data[56] were used to designate 1kb windows across the genome as either early (> 0.2) or late replicating (< -0.2) (Table 2.6). Again, the number of 1kb windows positive for either of these two replication timing states was tallied within larger 200kb windows to obtain vectors representing the density of early and late replicating regions of the genome, respectively. RNA-seq reads were binned in 200kb windows along the genome and the resulting read count totals were log-transformed to obtain a log-normal distribution more compatible with PCA. For gene density profiles, counts of unique transcription start sites from the UCSC MM9 refGene table within 200kb windows across the genome were obtained. This preprocessing step resulted in a 200 kb-resolution, 13,283 x 22 (windows x features) ESC feature matrix. This matrix of genomic feature data was passed to R's prcomp function[167]. Each column of data was scaled

prior to performing the PCA. We used the top three PCs that best characterized the genomic landscape of ESCs for further analysis (PC1/2/3) (Figures 2.6 and 2.8).

We investigated whether mappability was particularly associated with any of the principal components considered in the study by including the average mappability within 200 Kb windows in the input matrix for our PCA of genomic features. We found that mappability does not contribute significantly to any of the top three components discussed in our study (data not shown).

To describe the linear genomic state of ESCs and MEFs, features available for both cell types were selected for PCA (Table 2.6). Vectors containing feature counts within 200kb windows for each cell type were concatenated, allowing PCA to be conducted on the combined genomic features. The resulting 26,566 x 15 (concatenated windows x common features) ESC+MEF feature matrix was passed to R's prcomp function to conduct the PCA. Only PC1 was considered for the ES+MEF features (Figure 2.11A, 2.10A-D).

4.16 - Bait versus interactome comparisons based on PC score enrichment

To obtain a PC score enrichment value for each 4C bait, the mean PC score within the 200kb bait window and the four flanking windows (for a total of five 200kb bait windows = bait region) was calculated. The PC enrichment within the bait's interactome was calculated in two ways: 1) as a Spearman's *rho* statistic by correlating the vector of 4C hit percentages (average hit probabilities) within 200kb windows across the chromosome to the corresponding vector of PC scores; and 2) by calculating the mean PC score within 200kb windows that overlapped 4C

positive domains (as determined by the binomial test) by at least 25%. The five bait windows were excluded from the interactome enrichment. The rankings of the bait and interactome PC score enrichment values were then correlated to obtain the Spearman *rho* statistics shown in Figures 2.6E, 2.8C, and 2.8G.

4.17 - Curve fitting

We used R's loess[167] function to perform local fitting of a curve to PC data as a function of -$\log_{10}$(p-value) rank (Figure 2.13J-M, 2.12E, G-I), and subsequently R's predict.loess function to predict a loess fit and estimated standard error for each predicted value. An estimated 95% confidence interval was obtained by drawing a band +/- 2 s.e. on either side of the fitted curve.

4.18 - Hi-C comparisons

Mouse ESC, normalized, Hi-C interaction matrices based on 40kb bins were downloaded from the Ren Lab website (http://chromosome.sdsc.edu/mouse/hi-c/download.html)[42] and re-binned into 200kb bins to match the resolution of our 4C data. The row/columns within the Hi-C interaction matrix were extracted.

The rebinned Hi-C data (as described above) were also used for conducting "pseudo-4C", the Hi-C equivalents of the 4C-seq-based bait-interactome Spearman rank correlations in Figures 2.6E, 2.8C and 2.8G**.** Specifically, the row/columns within the Hi-C interaction matrix corresponding to each of the sixteen 4C baits analyzed in our ESC study were extracted. Bait PC score enrichments were calculated as described above for the 4C data sets by taking the mean PC score within the bait window and four flanking windows (five 200kb bait windows = bait region).

The mean PC score within the top and bottom 5% of windows based on read count in each row of the rebinned Hi-C matrices were similarly calculated. To ensure that the interactome enrichment was not driven by the strong contacts centered around the bait locus, we excluded regions around the bait locus ranging in size from 1 to 50Mb to show that the enrichment was robust across the length of the chromosome (Figure 2.4C/D, and data not shown). Furthermore, the vector of Hi-C read counts rebinned within the 200kb windows, besides the 5 bait windows, was Spearman rank correlated to the PC scores across the respective chromosome to obtain an enrichment value in terms of the *rho* statistic (Figure 2.6H, Table 2.5).

For the genome-wide bait-interactome PC score enrichment correlations (Figure 2.6G, 2.8D/H, 2.4C/D), each 200kb window of rebinned Hi-C data was treated as a pseudo-bait. Bait and interactome PC score enrichment values were calculated as described above for specific baits.

4.19 - Fluorescence *in situ* hybridization (FISH)

Cells were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM $MgCl_2$, 10mM PIPES, pH6.8), 30 seconds with CSK buffer containing 0.5% Triton X-100, and again for 30 seconds with CSK buffer, fixed in 4% PFA in 1xPBS for 10 minutes at room temperature, transferred to 70% ethanol, and stored at 4°C. Cells were dehydrated through a series 5 minute incubations in ice-cold 85%, 95%, and 100% ethanol, rehydrated in 2xSSC for 5 minutes, incubated in 2xSSC with 100ug/ml RNaseA (Invitrogen) at 37°C for 30 minutes, and washed three times for 5 minutes each in 2xSSC. DNA was denatured for 20 minutes at 80°C in 2xSSC with 70% deionized formamide (CalBiochem),

followed by immediate quenching in ice-cold 70% ethanol and a second dehydration series (performed as above). DNA FISH probes were denatured at 95°C for 5 minutes and allowed to pre-hybridize for 1 hour at 37°C before being added to dry slides. Probes were then allowed to hybridize with cellular DNA at 37°C for 16-48 hours in a humid chamber containing 50% formamide in 2x SSC. Following hybridization, cells were washed three times in each of the following solutions, pre-warmed to 42°C: 2x SSC/50% formamide, 2x SSC, and 1x SSC. The second 1x SSC wash contained 100 ng/ml DAPI to visualize nuclei. Slides were mounted in aquapolymount (Polysciences) and allowed to set overnight. FISH probes were generated from bacterial artificial chromosome (BAC) DNA (Figure 1c: *Pou5f1* locus – RP23-213M12; A – RP23-98F21; B – RP23-106C23; C – RP23-85E24; D – RP23-2B8; Figure S6c/d: *Hoxa* - RP24-283F1; *Hoxb* - RP23-290I2; *Hoxc* - RP23-473J19; *Hoxd* - RP24-398B4; *Sox2* - RP23-2B8) by incorporation of fluorescently labeled nucleotides (Cy3-dCTP, Perkin Elmer and Alexa 488-dUTP, Invitrogen) via biopriming (Invitrogen).

4.20 - 3D-FISH image acquisition and analysis

3D images were constructed from a series 0.2um z-stacks through selected individual ESCs or ESC colonies. 3D-distance measurements between FISH signal centers were acquired using the *Smart FISH3D* plugin for ImageJ[172,173]. Distance distribution statistics were calculated in R[167].

4.21 - Immunofluorescence and image analysis

V6.5 mouse ESCs were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM $MgCl_2$, 10mM PIPES, pH6.8), incubated for 1 hour with CSK buffer containing 0.7% Triton X-100, and washed again for 30

seconds with CSK buffer, fixed with 1xPBS/4% PFA for 10 minutes at room temperature, washed for 5 minutes in PBS/0.2% tween, incubated in blocking buffer (1x PBS, 10% goat serum, 0.2% fish skin gelatin, 0.2% tween) for 30 minutes, and incubated overnight in primary antibody diluted in blocking buffer at 4°C (anti-Nanog [eBioscience 14-5761-80]; anti-H3K27me3 [Active Motif 39155]; anti-RNAPII-S5 [Millipore 05-623]). Following primary antibody incubation cell were washed three times in PBS/0.2% tween and incubated in secondary antibody in blocking buffer. Cells were then wash three times in PBS/0.2% tween, with the second wash containing 1ug/mL DAPI, and mounted in Polyaquamount.

Image acquisition was performed on a spinning disc confocal microscope. Line intensities in Figure 5 were determined using Slidebook software from 3i. Images were exported as tif-files, and subsequent analysis was performed with customized R software[167]. Nuclear signal was identified via the removal of blue pixels whose intensities were below the 20th percentile of all blue pixel intensities for a given capture. Red and green signal outside of these preliminarily defined 'nuclear' signals were discarded. To remove nucleoli from the analysis 'nuclear' pixels with 0 values for either red or green pixels were similarly dropped from analysis. The remaining red and green pixels were quantile normalized, and the relationship between the remaining normalized red and green pixel positions and intensities were analyzed as shown in Figure 2.9**.**

4.22 - 3C

3C libraries were prepared in an analogous manner to 4C libraries, ending with the first ligation step. 3C primers are listed in Table 2.2, and PCRs were run with the following parameters: Primers A+B/C/D/E/F 95C 2 minutes, (95C 30 seconds, 60C 45 seconds, 72C 1 minute) x30,

72C 2 minutes, 1:5 dilution used as template for nested PCR with same parameters; Primers Z+V/X/Y/W - 95C 2 minutes, (95C 30 seconds, 68C 45 seconds, 72C 1 minute) x35, 72C 2 minutes.

FIGURES

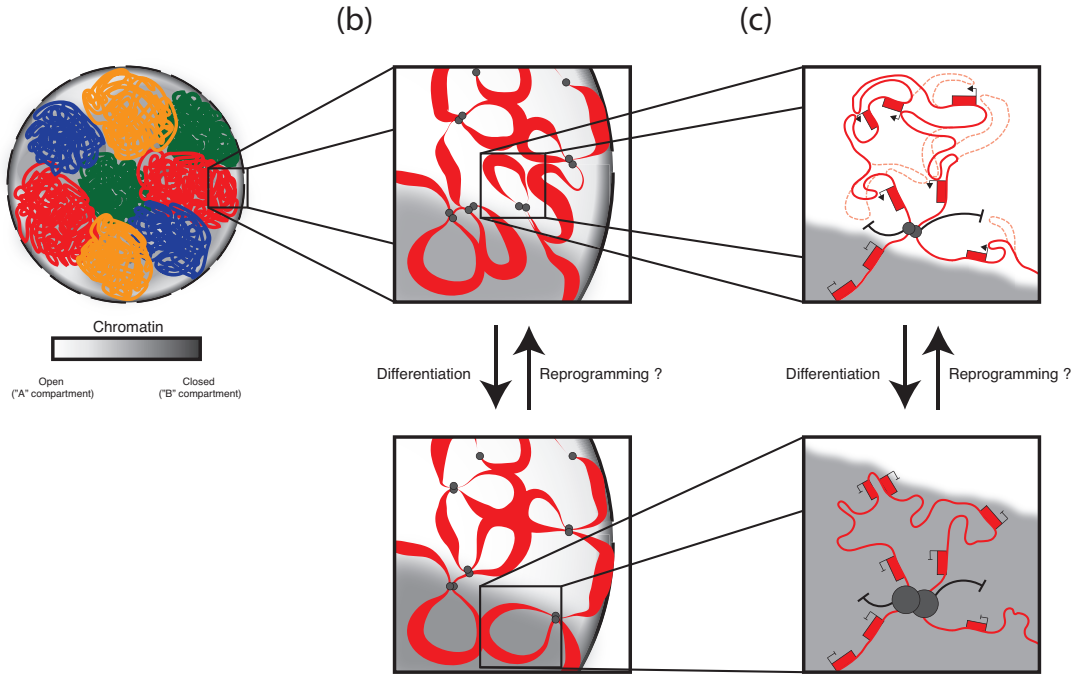

Figure 1.1 - Hierarchical levels of genome organization and their changes upon ESC differentiation

Model of the 3D organization of the genome in ESCs and its changes during the course of differentiation. We infer this model by combining findings from many different cell types. **(a)** Chromosomes exist as discrete, minimally overlapping territories. At the megabase level, compartments of open (white) and closed (gray) chromatin coarsely divide the genome into regions enriched for features of euchromatin and those depleted of euchromatic features respectively. Locus positioning within the open or closed compartment defines likely interaction partners, with loci in the open compartment interacting more frequently with others in open, and those in closed more frequently with others closed. Loci potentially switch compartments during differentiation, somatic cell reprograming, or between different cell types as their euchromatic character changes. **(b)** Below the megabase, level the genome is divided into topological domains, which, we speculate, exist as large chromatin loops created by the juxtaposition of CTCF binding sites at TAD boundaries. These TADs function as modular units of genome organization whose member genes are often co-regulated, and we propose, localize as units within the nucleus. **(c)** Within TADs, enhancers and promoters loop together extensively and

promiscuously to orchestrate cell type-specific gene expression profiles, with co-regulation of intra-TAD genes occurring both during gene expression and transcriptional silencing.

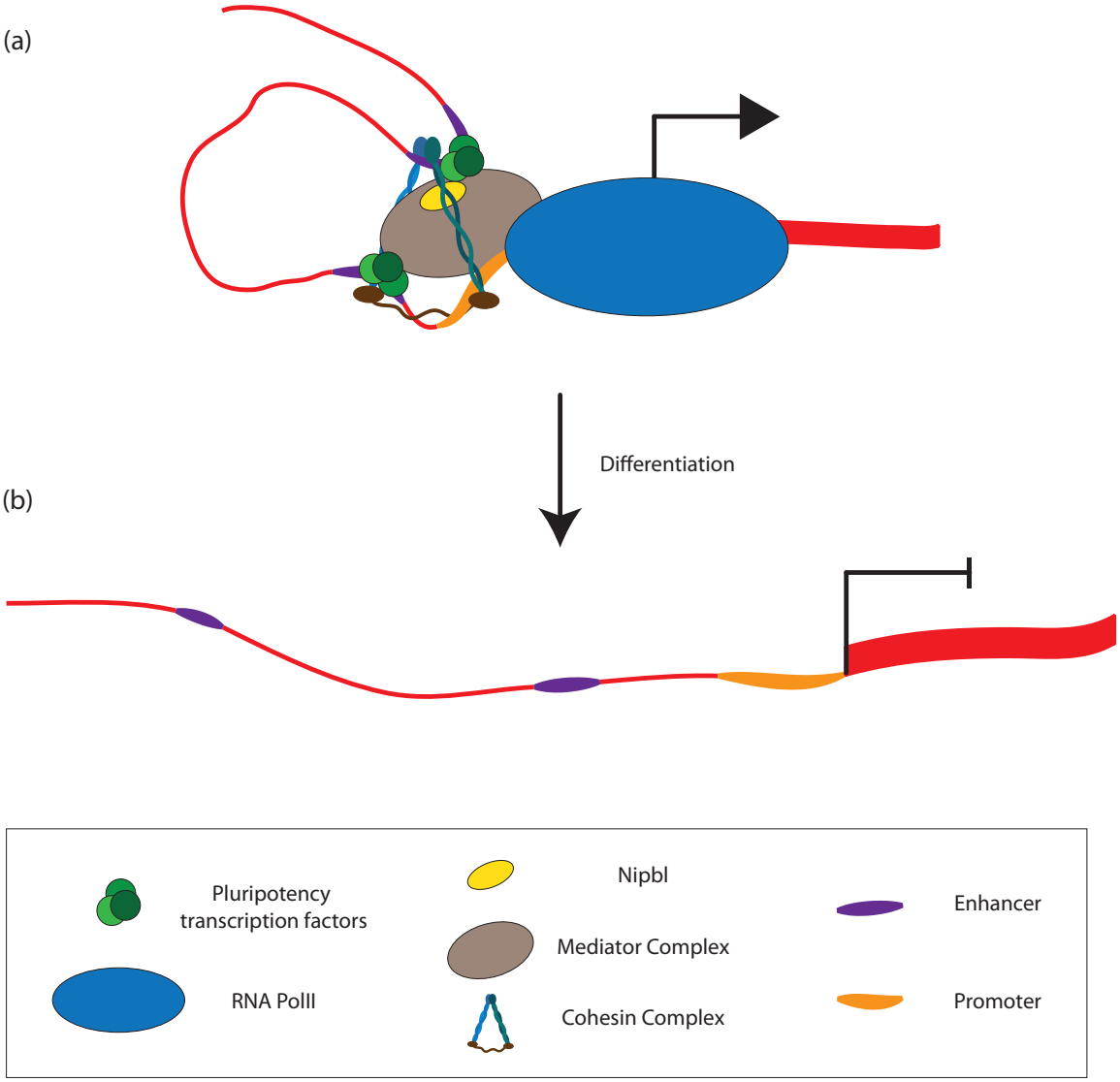

Figure 1.2 - The Mediator complex recruits Cohesin to chromatin and facilitates cell-type specific enhancer-promoter looping and gene expression

Inspired by the work of Kagey et al. and others referenced in the text, **(a)** Mediator-recruited Cohesin complexes orchestrate cell-type specific enhancer-promoter loops, providing a mechanism for the cell-type specific action of an enhancer on a given promoter and, by extension, cell-type specific gene expression patters. In ESCs, many Oct4 (O), Sox2 (S), and Nanog (N) bound regions of the genome coincide with Mediator and Cohesin occupancy. **(b)** In fibroblasts, where OSN binding is absent, Mediator and Cohesin show differential DNA binding patters and enhancer-promoter looping events at many ESC-specific gene loci are absent.

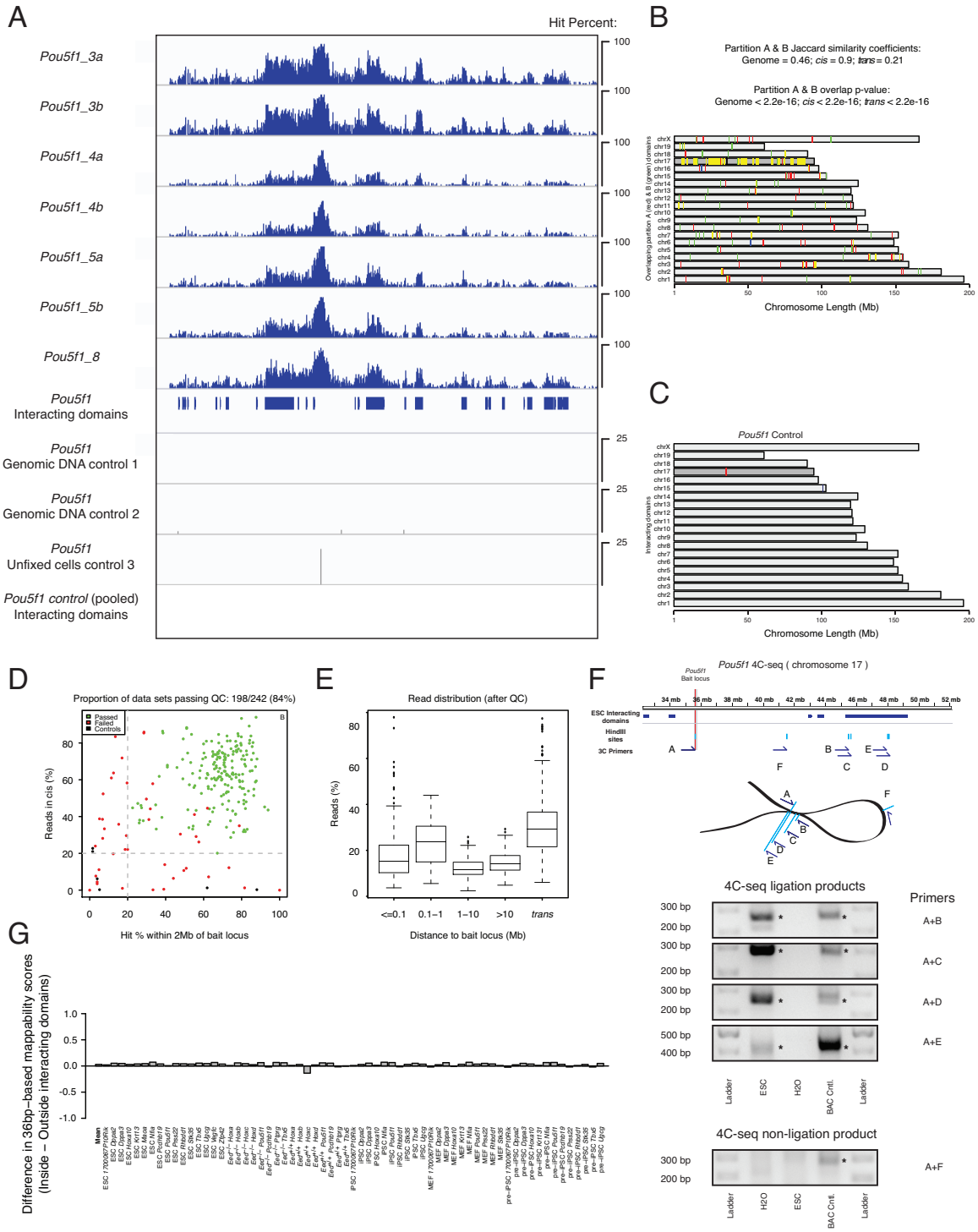

Figure 2.1 – Reproducibility, validation, and quality control of 4C-seq data

A, Integrative Genomics Viewer tracks demonstrating the reproducibility of *Pou5f1* 4C-seq hit percentage data for four biological and three technical replicate data sets. Biological replicates are designated with numbers, technical replicates with letters. The interacting domains as

identified by the binomial test for the pooled 4C-seq *Pou5f1* data are given below the hit percentage tracks. 4C-seq experiments with genomic DNA and control libraries from unfixed ESCs follow, demonstrating the lack of significant interactions.

B, To determine the reproducibility of our data in both *cis* and *trans*, *Pou5f1* replicates were pooled (blue) or partitioned into two groups (A and B, red and green, respectively) and interacting regions were determined for each set and are shown on the genome-wide plot. Yellow regions are those that were called as interacting domains in both A and B partitions, and blue regions were called as interacting domains in only the pooled data set. Jaccard similarity coefficients are noted for the overlap of significant domains between A and B, genome wide, *cis* only, and *trans* only, and the significance of the overlap was determined by the hypergeometric test. The darker chromosome represents the *cis* chromosome.

C, 4C-seq experiments on pooled control libraries, here shown for the *Pou5f1* locus (*Pou5f1* Control) in ESCs, display no significant interactions genome-wide. The red mark denotes the bait locus, and the darker chromosome represents the *cis* chromosome.

D, Quality control (QC) metrics. All 4C-seq data sets used in this study were required to have at least 20% of HindIII sites covered by reads within 2Mb of the bait locus (x-axis), as well as at least 20% of their reads occurring in *cis* to the bait locus (y-axis), among other quality control metrics (see *Experimental Procedures*). Only libraries that passed these QC metrics were analyzed further and are included in this study.

E, Read distribution as a function of distance from the bait locus for all data sets that passed QC. This analysis excludes the self-ligated and undigested products.

F, 3C validation of *Pouf51* interactions. (i) Schematic representation of the 3C experimental design to confirm interactions between the *Pou5f1* locus (primer A) and distal regions of the same chromosome (in *cis*), including the four interacting fragments (primers B-E) and an intervening non-interacting fragment (primer F). (ii) 3C PCR results for the setup described in (i) confirming the presence of ligation products resulting from the juxtaposition of the genomic regions (B-E) with A, and an absence of ligation products between A and the intervening genomic region (F). An H_2O control and a BAC-generated positive control are shown for each PCR product. The asterisks mark the respective PCR products.

G, UCSC ENCODE mm9 36mer-based mappability scores (ftp://encodeftp.cse.ucsc.edu/pipeline/mm9/wgEncodeMapabilitywgEncodeCrgMapabilityAlign36mer.bigWig) were

compared between the interacting domains (determined based on our 4C-seq analysis pipeline) and the scores outside these domains. The graph displays the difference in the mean 36-mer-based mappability scores inside and outside of the intrachromosomal interaction domains for the indicated baits and cell types, as well as the average across all baits and cell types (labeled 'MEAN'), demonstrating that our 4C-seq results are not biased by mappability.

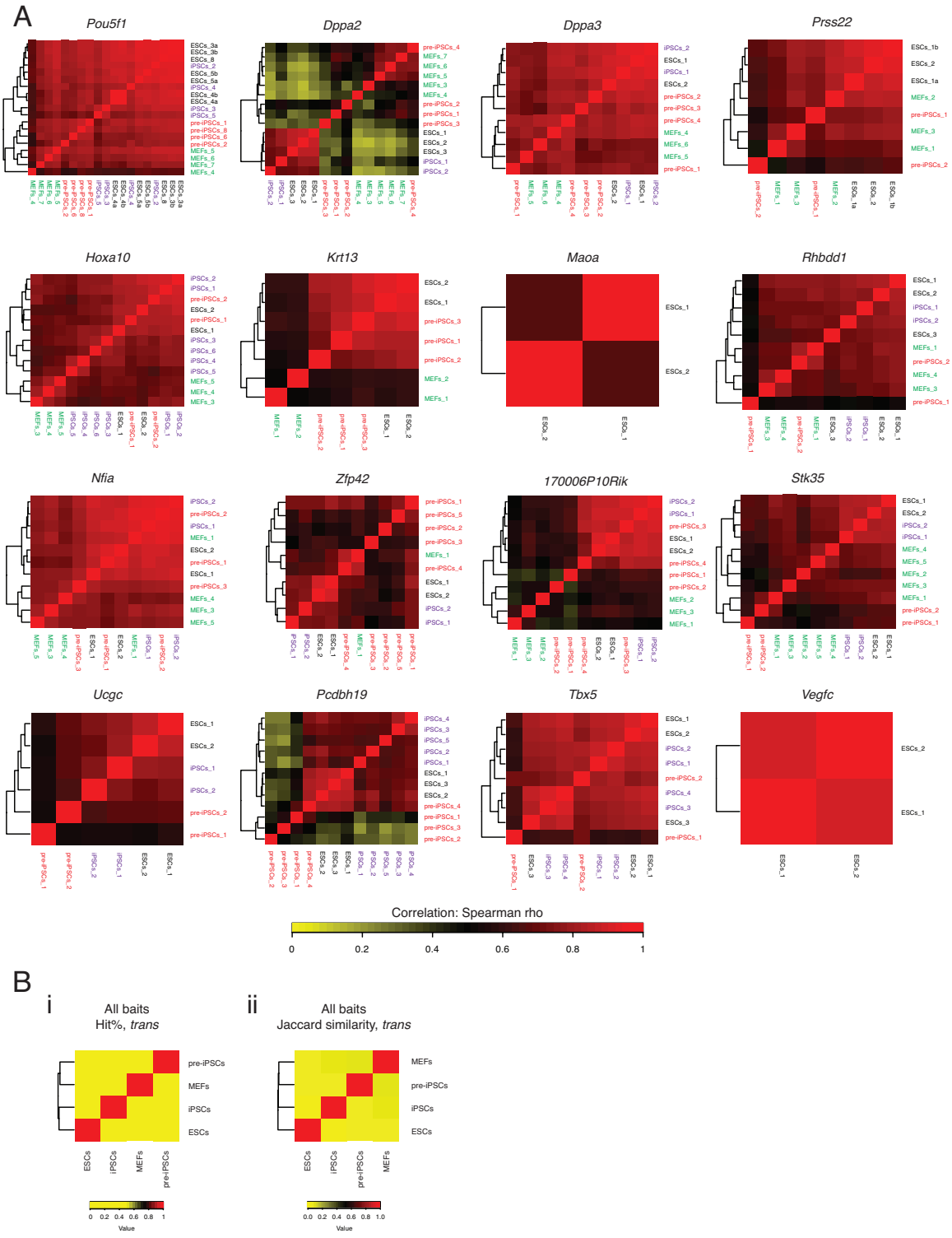

Figure 2.2 - 4C-seq replicate data sets cluster by cell type, revealing pluripotency-specific chromatin contacts

A, Spearman rho correlation matrixes based on the percent of HindIII sites hit within 1000kb windows in *cis* for individual 4C replicates in wildtype ESCs (V6.5 line), iPSCs, pre-iPSCs, and

MEFs for the noted baits. Note that the preferential clustering of individual replicates by cell type, and the high correlations of the data between ESCs and iPSCs. Correlation values are indicated by the key, and numbers listed next to cell types correspond to biological replicates, while technical replicates are designated with letters.

B, (i) Unsupervised hierarchical clustering of Spearman rank correlation values of the hit percentages within 200kb windows along the *trans* chromosomes of eight different bait loci (*Pou5f1*, *Stk35*, *1700067P10Rik*, *Nfia*, *Dppa3*, *Rhbdd1*, *Hoxa10,* and *Dppa2*) in wildtype ESCs (V6.5 line), MEFs, iPSCs, and pre-iPSCs, demonstrating the pluripotency-specific organization of interchromosomal chromatin contacts within the mouse genome. Color key defines Spearman *rho* values. (ii) Unsupervised hierarchical clustering of Jaccard similarity coefficients for the overlapping interacting domains *trans* in ESCs, iPSCs, pre-iPSCs, and MEFs, for the same eight bait loci as (A). Color key defines Jaccard similarity values.

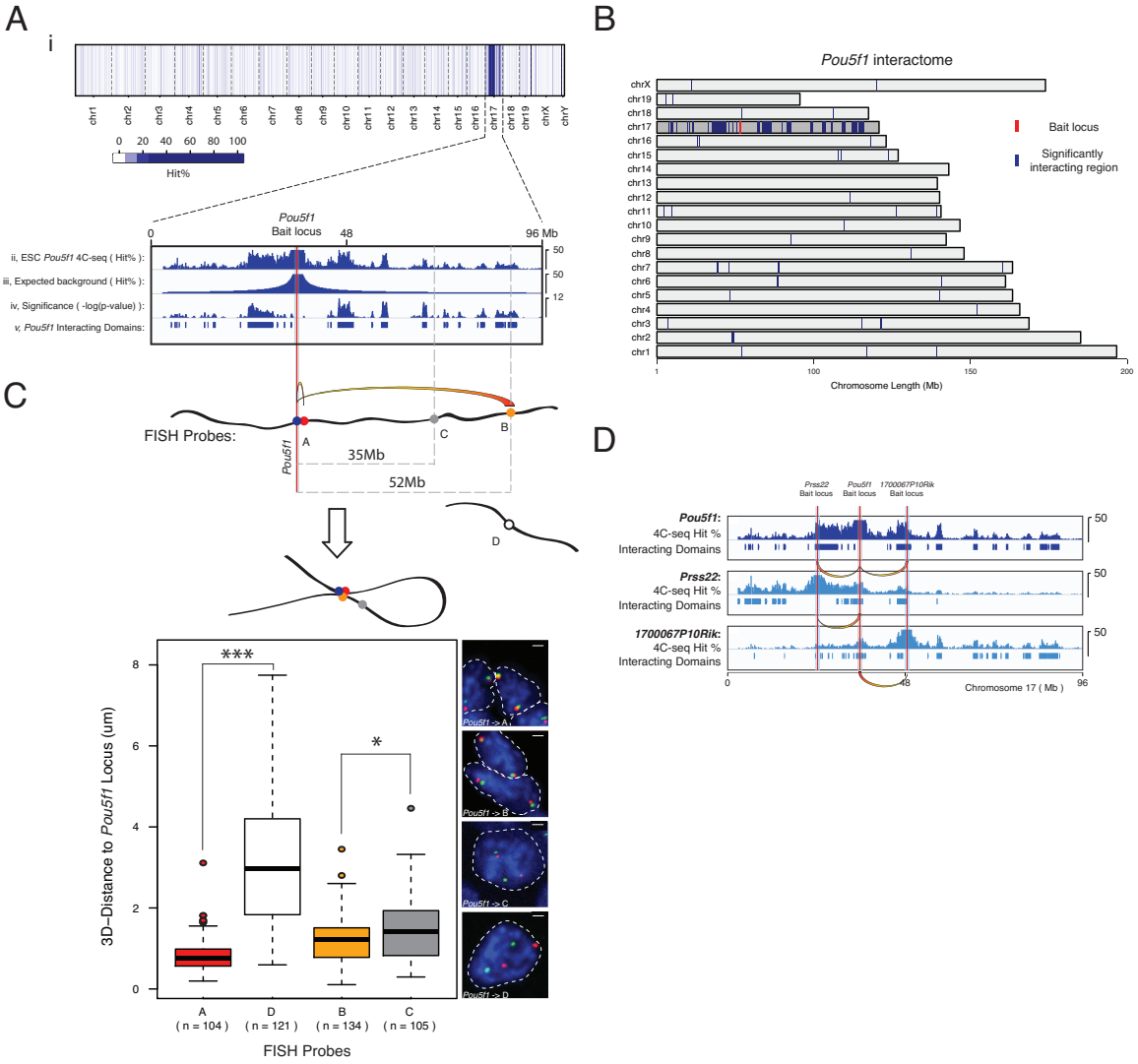

Figure 2.3 – Long-range chromatin contacts of the *Pou5f1* bait region in ESCs

A, Results from a representative, pooled 4C-seq data set in mouse ESCs, depicting the genome-wide interaction profile of the *Pou5f1* locus on chr17 (*cis* chromosome). (i) Average hit probability (hit percentage) within 200kb windows tiled across the entire genome. (ii) The hit percentage in *cis* (cut off at the maximal value displayed) was compared to an empirically modeled background hit probability (iii) using the binomial test. The binomial test result (-log(p-values); (iv)) and the resulting significantly interacting domains of the *Pou5f1* locus ((v)**;** based on -log(p-values) >= 1.8) are shown. The vertical red line and associated gray bar denote the *Pou5f1* bait locus and the extended 1Mb bait region, respectively, which is excluded from any downstream analysis of the *Pou5f1* interactome.

B, Visualization of the significantly interacting domains (blue) of the *Pou5f1* bait locus (red) genome-wide.

C, DNA FISH was performed between the *Pou5f1* locus and genomic regions marked by A-C (in *cis*) or by D (in *trans*), and 3D distances were measured. Regions A and B were found to interact with the *Pou5f1* locus based on 4C-seq, while the region "C" was non-interacting and is located in an intervening genomic region closer to *Pou5f1* on the linear DNA than "B". * = p-value < 0.05, *** = p-value < 0.001, Wilcoxon rank-sum test.

D, Reciprocal 4C-seq experiments confirm interactions of the *Pou5f1* locus. Two genomic regions containing the gene *Prss22* and *1700067P10Rik* were identified as interacting partners of the *Pou5f1* locus. Reciprocal 4C-seq experiments using *Prss22* and *1700067P10Rik* as bait loci confirmed their interaction with *Pou5f1*. The bait regions, 4C-seq average hit probability (hit percentage) within 200kb windows, and the interacting domains are depicted for the three 4C-seq data sets.

Figure 2.4

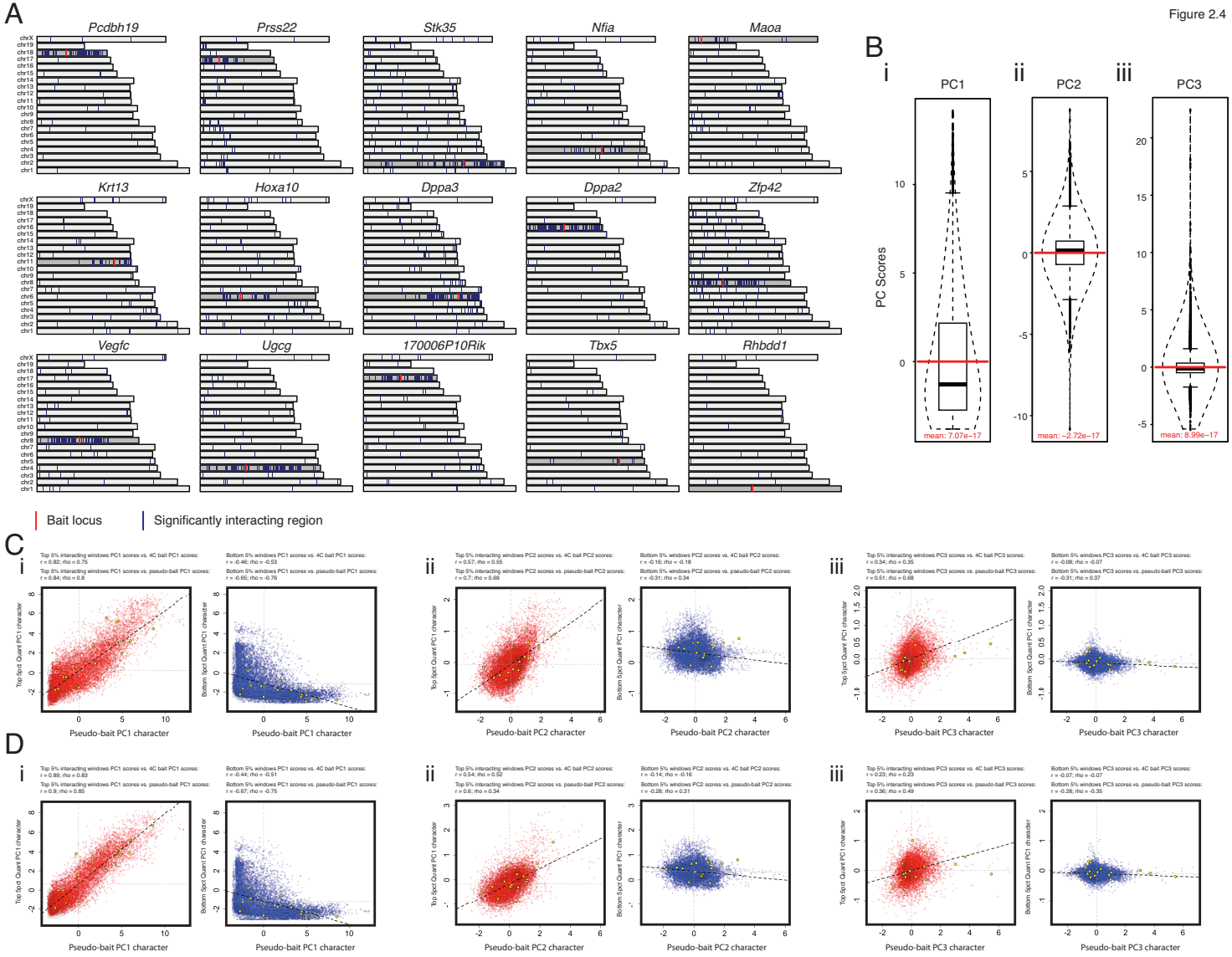
A
Pcdbh19
Prss22
Stk35
Nfia
Maoa
Krt13
Hoxa10
Dppa3
Dppa2
Zfp42
Vegfc
Ugcg
170006P10Rik
Tbx5
Rhbdd1
Bait locus
Significantly interacting region
B
i
PC1
ii
PC2
iii
PC3
PC Scores
mean: 7.07e−17
mean: −2.72e−17
mean: 8.99e−17
C
i
ii
iii
Pseudo-bait PC1 character
Pseudo-bait PC2 character
Pseudo-bait PC3 character
D
i
ii
iii
Pseudo-bait PC1 character
Pseudo-bait PC2 character
Pseudo-bait PC3 character

Figure 2.4 – Genome-wide analysis of ESC interactomes

A, We obtained 4C-seq data for a total of 16 baits in the V6.5 ESC line. The interacting domains for the Pou5f1 4C-seq data set are shown in Figure 1, and the interacting domains identified for the remaining 15 baits are depicted here genome-wide, with their relative chromosomal locations. Blue marks represent significantly interacting windows, red marks the bait locus, and the darker chromosome denotes the *cis* chromosome.

B, Demonstration of the continuous nature of PC scores. (i) Box plot of genome-wide PC1 scores overlaid with a violin plot. (ii) As in (i), but for PC2. (iii) As in (i), but for PC3.

C, Hi-C data analysis demonstrates that the preferential co-localization of genomic regions with similar genomic features identified by 4C-seq is a genome wide phenomenon for interactions in *cis*: Pseudo-4C analysis was performed on Hi-C data as described in Figure 2.6E/F. (i) (left) The mean PC1 score within the top 5% of 200kb windows (ranked by read count based on Hi-C data) along the cis-chromosome, excluding the 1Mb bait region, and the mean PC1 score within the respective 1Mb pseudo-bait region were determined, and plotted as red point on the scatterplot. If the bait region is one of the our baits analyzed by 4C-seq in Figure 2.6E-i, the data point was plotted in yellow (4C-bait loci). Correlations between bait and interactome PC1 scores are noted. The data are also summarized by the regression line in black, and the mean bait and interactome PC1 scores are demarcated by vertical and horizontal grey lines, respectively, and contour lines represent data density. The same data are shown in Figure 2.6G. (i) (right) Similar to (i), an analysis was performed for the 5% of windows that are the least likely to interact based on Hi-C read count (bottom 5%) and their mean PC1 score was plotted in blue against the mean score of the 1Mb bait regions. Again, if the bait region is one of the our baits analyzed by 4C-seq in Figure 2.6E-i, the data point was plotted in yellow (4C-bait loci). (ii) As in (i), except that the analysis was performed for PC2 scores, and the red scatterplot is repeated from Figure 2.8D. (iii) As in i,), except that the analysis was performed for PC3 scores, and the red scatterplot is repeated from Figure 2.8H.

D, As in (C), except that all 200kb windows within the 10Mb region around the bait loci were excluded from the determination of the mean PC scores within the top and bottom 5% of 200kb windows along the cis-chromosome. Importantly, the correlation values between interactome and bait character, with regards to PC scores, were not significantly changed upon exclusion of the 10Mb region around the bait loci, or of increasingly larger proximal interactions around the bait (up to 25Mb on either side of the bait region, data not shown), indicating that even distal interactions follow the logic that regions with similar PC character interact.

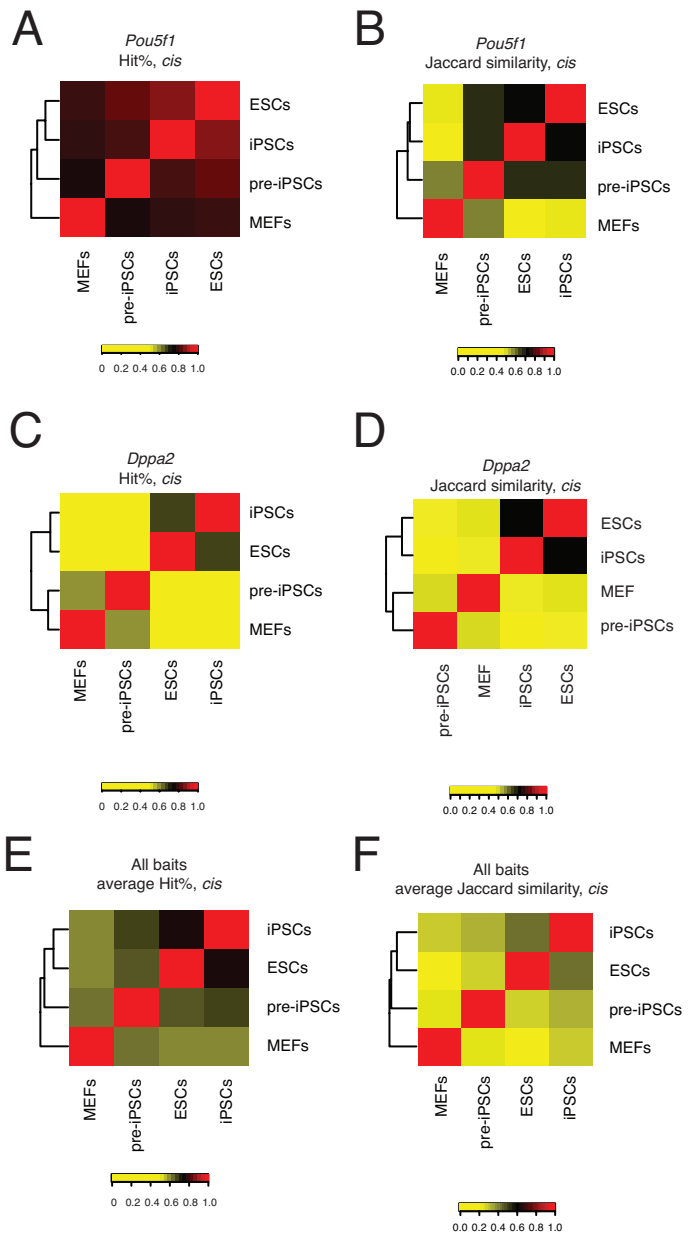

Figure 2.5 – Long-range chromatin contacts change upon differentiation of ESCs and are reset to an ESC-like state upon reprogramming of somatic cells to iPSCs

A, Unsupervised hierarchical clustering of Spearman rank correlation values of the hit percentages within 200kb windows along the *cis* chromosome (chr17) for *Pou5f1* 4C-seq-derived interactomes in ESCs, iPSCs, pre-iPSCs, and MEFs. Color key defines the Spearman *rho* values.

B, Unsupervised hierarchical clustering of Jaccard similarity coefficients for the overlap of *Pou5f1* interacting domains in *cis* between ESCs, iPSCs, pre-iPSCs, and MEFs. Color key defines the Jaccard similarity values.

C, As in (A), except for the *Dppa2* bait locus.

D, As in (B), except for the *Dppa2* bait locus.

E, As in (A), except across eight different bait loci in ESCs, MEFs, iPSCs, and pre-iPSCs (*Pou5f1*, *Stk35*, *1700067P10Rik*, *Nfia*, *Dppa3*, *Rhbdd1*, *Hoxa10,* and *Dppa2*).

F, Unsupervised hierarchical clustering of the average Jaccard similarity coefficients for the overlap of interacting domains *cis* in ESCs, iPSCs, pre-iPSCs, and MEFs across the same eight bait loci as in (E).

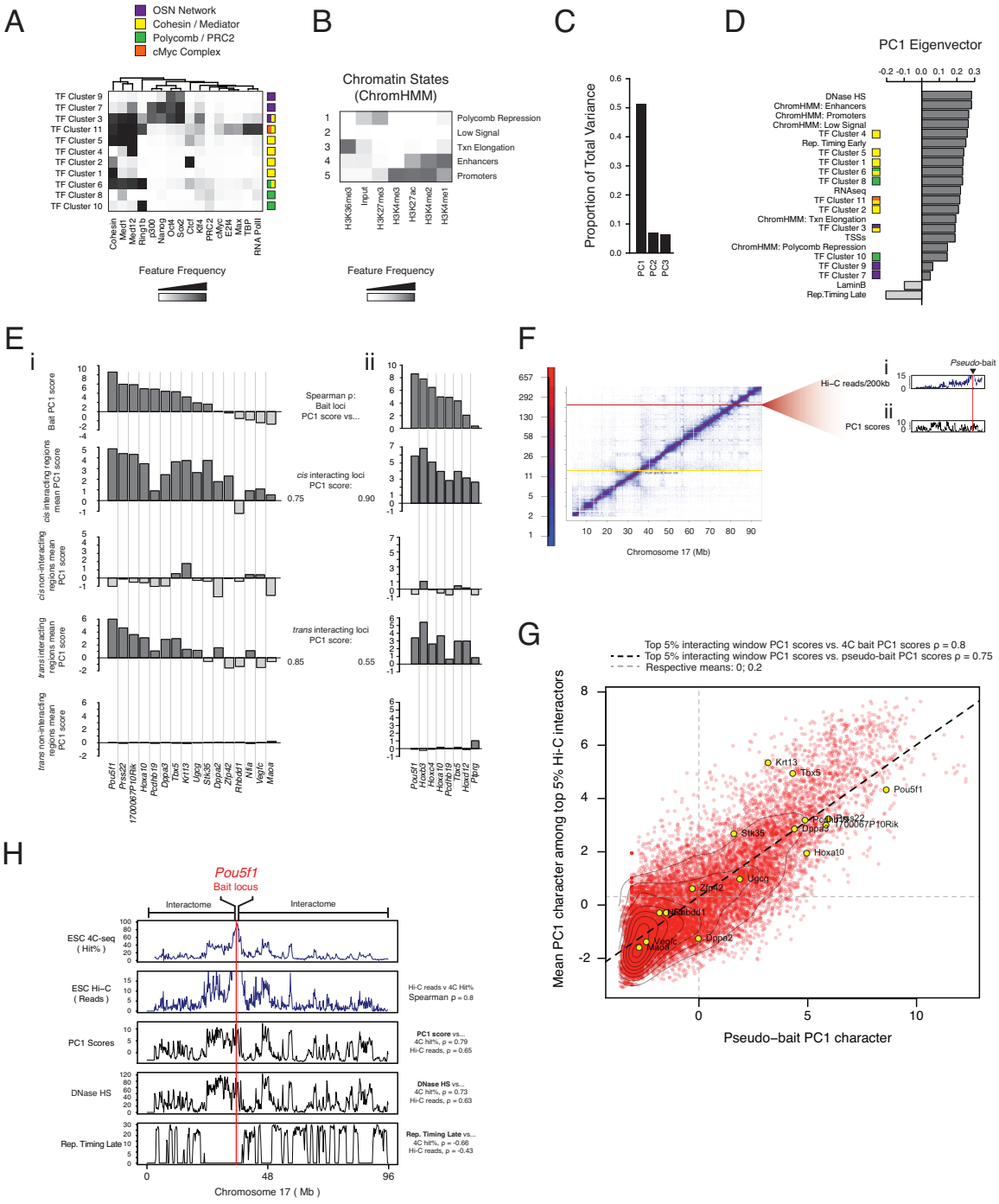

Figure 2.6 – Interactions between regions with similar open/closed chromatin states are an intrinsic aspect of chromosome conformation

A, Transcription factor (TF) clusters were defined using k-means clustering at 1kb resolution with noted data sets and annotated based on the proportion of windows within the cluster bound by each feature as represented by the grey scale of the heatmap (feature frequency, see

Experimental Procedures). The cluster annotations reflect known transcriptional networks in ESCs as well as extensive enrichment for Cohesin and Mediator (Med1 and Med12) binding.

B, Chromatin states were determined by a multivariate hidden Markov model, at 200bp resolution [123]. The plot shows the emission parameters learned *de novo* on the basis of genome-wide recurrent combinations of the indicated histone marks. Grey scale denotes the frequency with which a given mark is found at genomic positions corresponding to the chromatin state.

C, PCA was performed on the genomic features of ESCs including the chromatin states and TF clusters from (A) and (B), as well as RNA-seq expression data, Dnase hypersensitivity, LaminB binding, early and late DNA replication timing (Rep. timing), and density of transcriptional state sites (TSSs). The graph depicts the proportion of total variance in genomic features described by each principal component.

D, PC1 eigenvector ranked by genomic feature contribution.

E, (i) Top to bottom: Mean PC1 score for the 1Mb region centered on each listed bait's locus; mean PC1 score within interacting regions in *cis*; mean PC1 score within non-interacting regions in *cis*; mean PC1 score within interacting regions in *trans*; mean PC1 score within non-interacting regions in *trans*. Spearman rho's give the correlation between the mean PC1 score of the bait regions and interactomes across all analyzed baits in *cis* and *trans*. (ii) Identical analysis to (i) for the independently derived $Eed^{+/+}$ ESC line discussed in Figure 7 with a partially overlapping set of bait loci.

F, Schematic of genome-wide pseudo-4C analysis based on Hi-C data. We extracted each row of the Hi-C contact matrix, adapted from Dixon et al.[42], which represents the interactome of one pseudo-bait (i) (see *Experimental Procedures*), and analyzed the PC1 character of the bait and interactome (ii). Specifically, the mean PC1 score within the top 5% 200kb windows (ranked by reads and excluding 1Mb around the pseudo bait) and the mean PC1 score within the respective 1Mb pseudo-bait region (Pseudo-bait PC1 character) were determined, and plotted as a point on the scatterplot shown in (G). These points are colored in red (pseudo-baits), but if the bait region is one of the our baits analyzed by 4C-seq in (E), the data are plotted in yellow (4C-bait loci).

G, Result of the analysis described in (F). Bait loci analyzed by 4C-seq (E, yellow points with labels) show a similar trend when analyzed based on Hi-C data. Overall correlations between bait and interactome PC1 scores are noted. The Hi-C data are also summarized by the regression

line in black, the mean bait and interactome PC1 scores are demarcated by vertical and horizontal grey lines, respectively, and contour lines represent data density.

H, Comparison, from top to bottom, of the *Pou5f1* 4C-seq-based *cis*-interactome (defined here by the average hit percentage within tiled 200kb windows across chr17), the *Pou5f1 cis*-interactome defined by Hi-C read counts, PC1 scores, DNase HS, or late DNA replication timing along chromosome 17. Correlation values are specified. The *Pou5f1* bait locus is indicated with a red line and associated grey bar. The data demonstrate that the *Pou5f1* intrachromosomal interactome correlates more strongly with PC1 than with individual genomic features previously shown to correlate with chromatin organization.

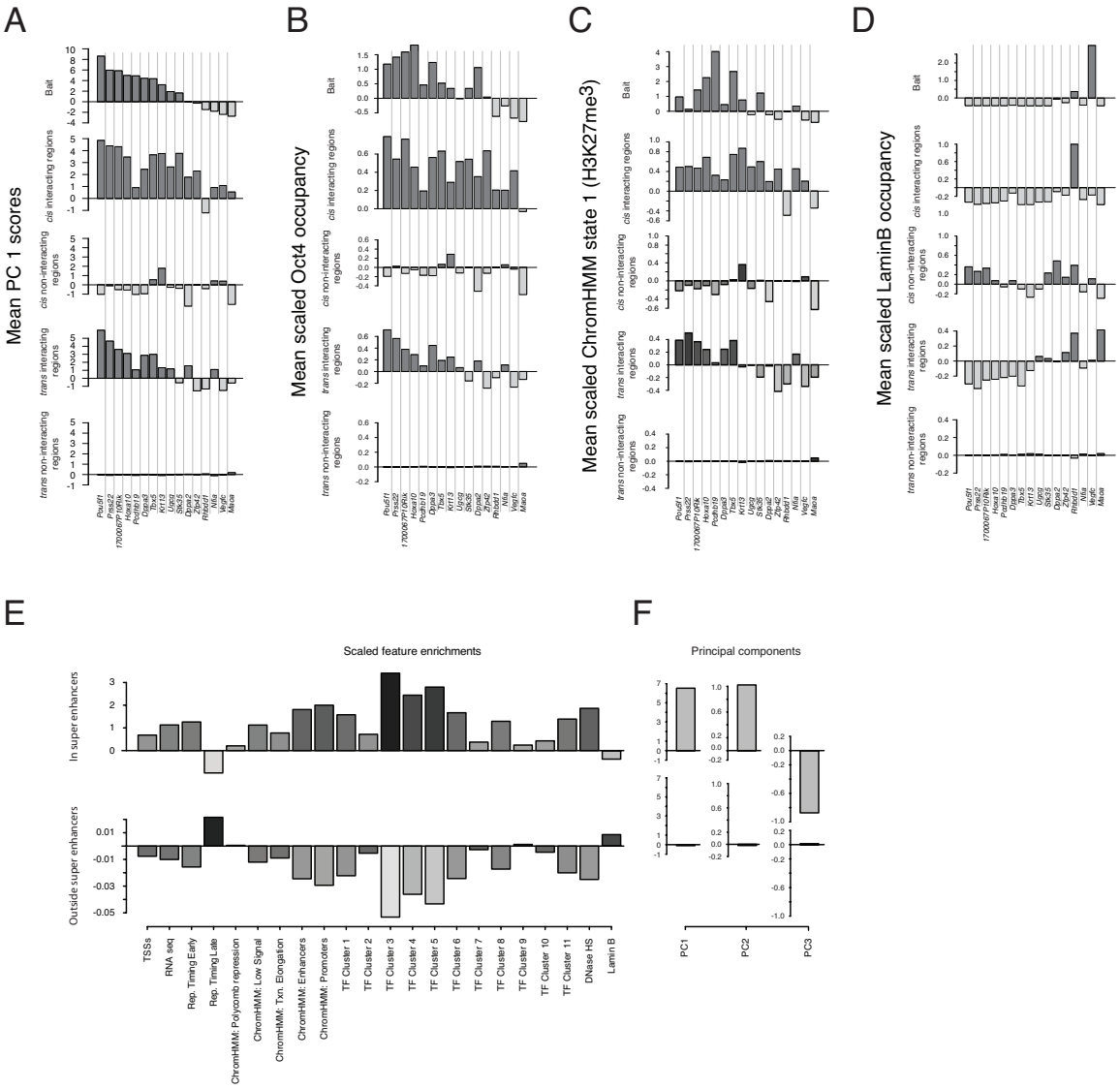

Figure 2.7 – Relationship between PC1 and individual feature enrichment

A-D, Comparison of PC1 and feature scores between bait regions and their interactomes: **A,** Top to bottom: Mean PC1 score for the 1Mb region centered on each bait locus in ESCs; mean PC1 score within interacting regions in *cis* for each bait locus; mean PC1 score within non-interacting regions in *cis* for each bait locus; mean PC1 score within interacting regions in *trans* for each bait locus; mean PC1 score within non-interacting regions in *trans* for each bait locus. Spearman rho's give the correlation between PC1 bait character and interactome character across all analyzed baits in *cis* and *trans*. These results are duplicated from Figure 3E-i for illustration. **B,** As in (A), but showing the average, standardized Oct4 enrichment within the bait regions, and their interacting and non-interacting regions. **C,** As in (B), but for H3K27me3 (which is captured

by ChromHMM state 1). **D,** As in (B), but for LaminB enrichment. The data shown in (A-D) suggest that due to the strong enrichment of transcription factors, chromatin regulators, transcriptional machinery, Cohesin, and active chromatin states in genomic regions with high PC1 scores and their preferential co-localization in 3D space, their interactomes are also highly enriched for any of the features positively correlating with PC1. Indeed, the extent of Oct4 and H3K27me3 enrichment in a bait's interacting regions mirrors its PC1 character. Conversely, LaminB occupancy, which correlates with closed, PC1 negative chromatin character, shows the opposite relationship. As such, the enrichment of features positively correlating with PC1 in a given bait's interactome may not be indicative of specific mechanistic roles for each enriched feature but may be instead a consequence of the overarching nature of chromosomal conformation.

E, Relationship of ESC super enhancers to the genomic features tested in our study. Enrichment of chromatin states and transcription factor clusters within and outside of ESC super-enhancers.

F, Mean PC1, PC2, and PC3 scores within and outside of ESC super-enhancers.

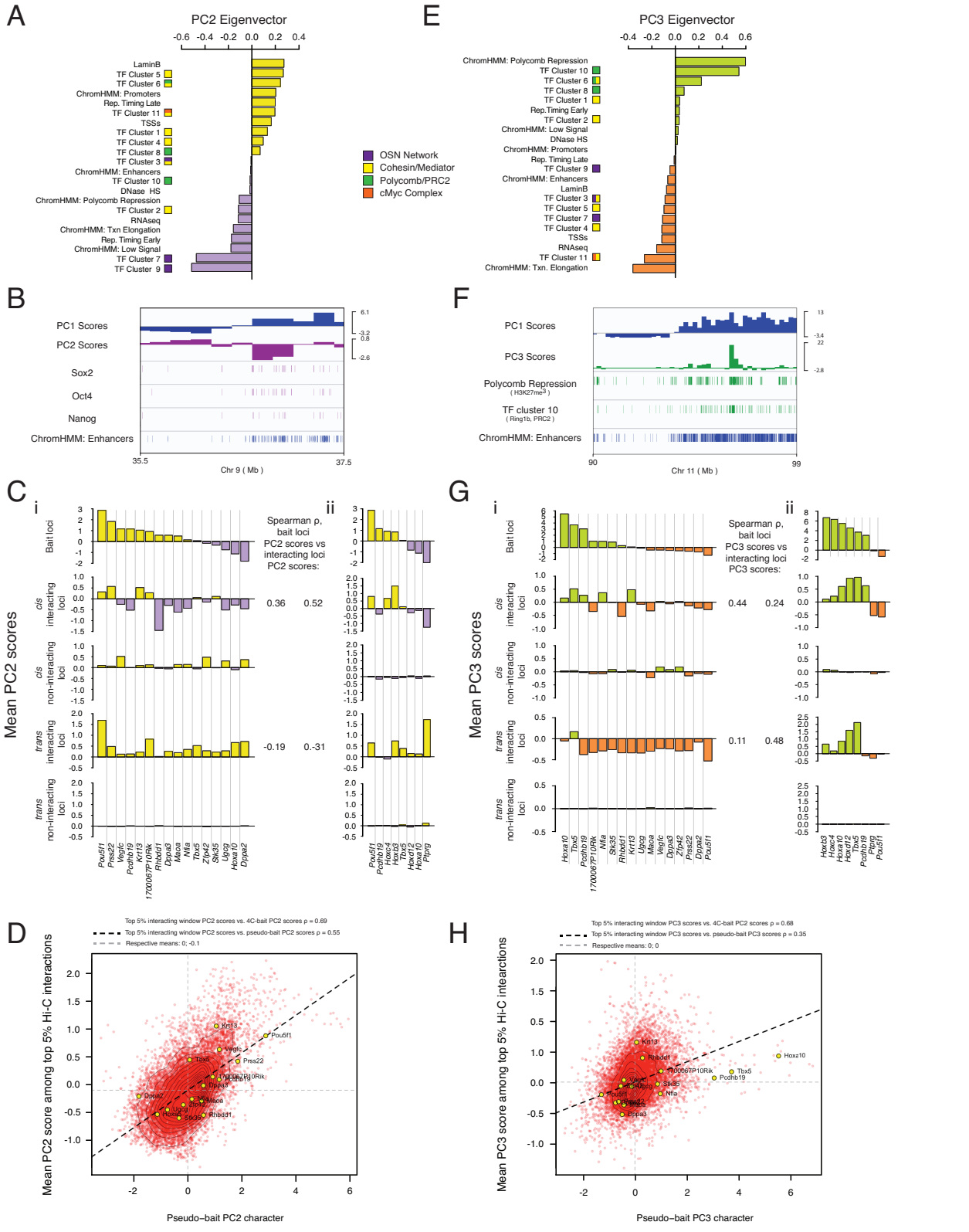

Figure 2.8 – Chromatin interactions preferentially occur between regions with shared chromatin states and ESC-specific transcriptional networks

A, PC2 eigenvector with individual feature contributions. TF clusters and chromatin states are defined in Figure 3.

B, Integrative Genomics Viewer tracks showing a representative genomic region with PC1 and PC2 scores; Sox2, Oct4, and Nanog occupancy; as well as enhancer density.

C, (i) Top to bottom: Mean PC2 score for the 1Mb region centered on each bait's locus; mean PC2 score within interacting regions in *cis*; mean PC2 score within non-interacting regions in *cis*; mean PC2 score within interacting regions in *trans*; mean PC2 score within non-interacting regions in *trans*. Spearman rho's give the correlation between the baits' PC2 and the interactomes' PC2 character across all analyzed baits in *cis* and *trans*. (ii) Identical analysis to (i) except for an independently derived $Eed^{+/+}$ ESC line discussed in Figure 2.13 with a partially overlapping set of bait loci.

D, Genome-wide pseudo-4C analysis of Hi-C data as described in Figure 2.6F/G, except for PC2.

E, PC3 eigenvector with individual feature contributions.

F, Integrative Genomics Viewer tracks showing a representative genomic region with PC1 and PC3 scores; H3K27me3 and TF cluster 10 (Ring1b/PRC2) enrichment, as well as enhancer density.

G, As in (c), except for PC3 scores.

H, As in (d), except for PC3 scores.

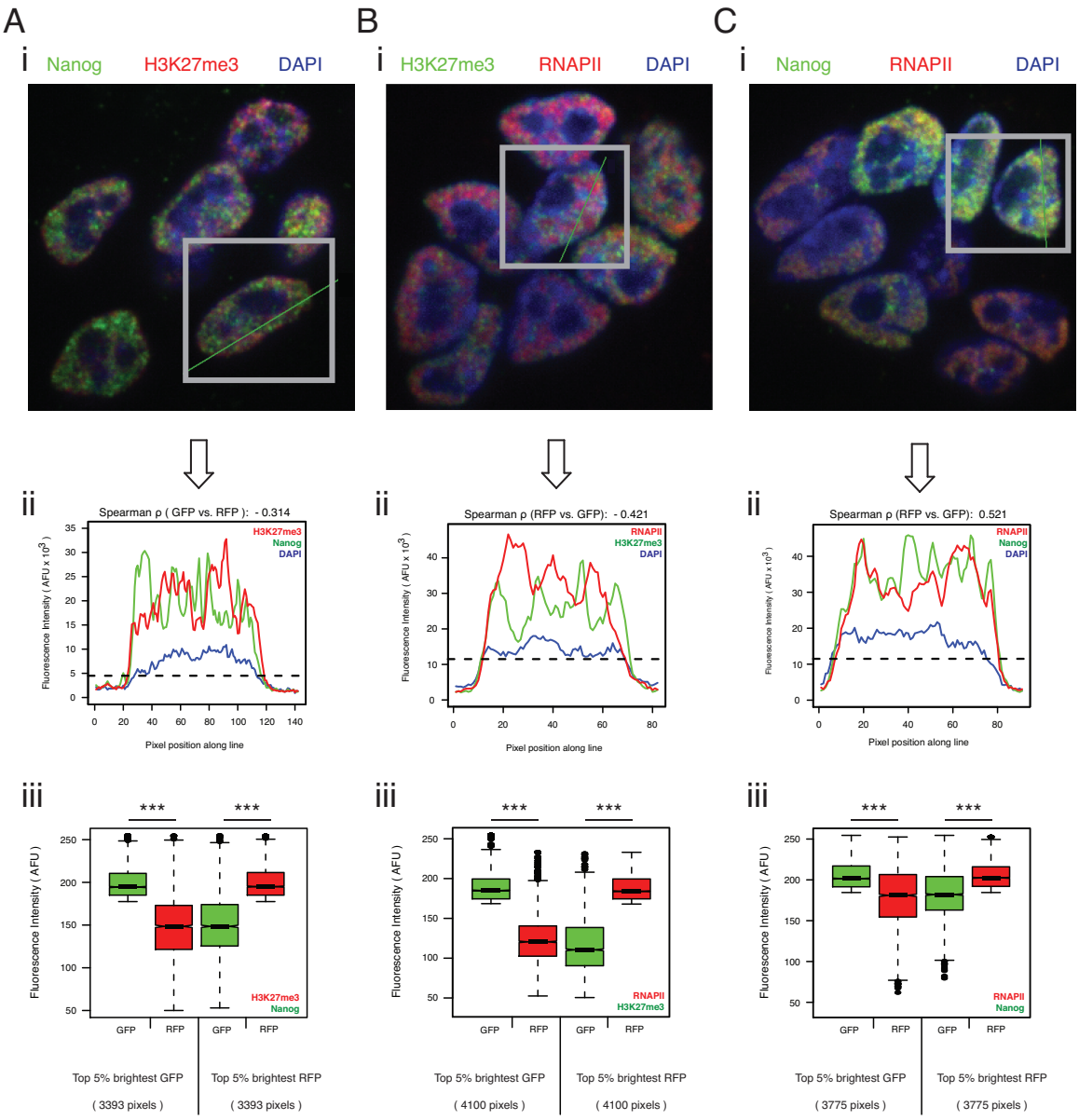

Figure 2.9 - Nanog and H3K27me3 segregate in the ESC nucleus

A, (i) Image of ESCs immunostained with antibodies against Nanog (green) and H3K27me3 (red). Nuclei are stained with DAPI (blue). (ii) Red and green pixel intensities along the measurement line in (i) for all pixels whose DAPI signals were above the indicated threshold (dotted line). (iii) Quantile normalized fluorescence intensity distribution of the top 5% brightest nuclear (but non-nucleolar) green pixels (GFP) and the normalized red pixel intensity (RFP) at the corresponding position (left); and normalized fluorescence intensity distribution of the top 5% brightest nuclear (but non-nucleolar) red pixels and the normalized distribution of green

pixel intensity at the corresponding position (right), for all cells in the ESC colony shown. *** = p-value < 2E-16, Wilcoxon rank-sum test.

B, As in (A), except for ESCs stained with antibodies against H3K27me3 (green) and RNAPII (red).

C, As in (A), except for ESCs stained with antibodies against Nanog (green) and RNAPII (red).

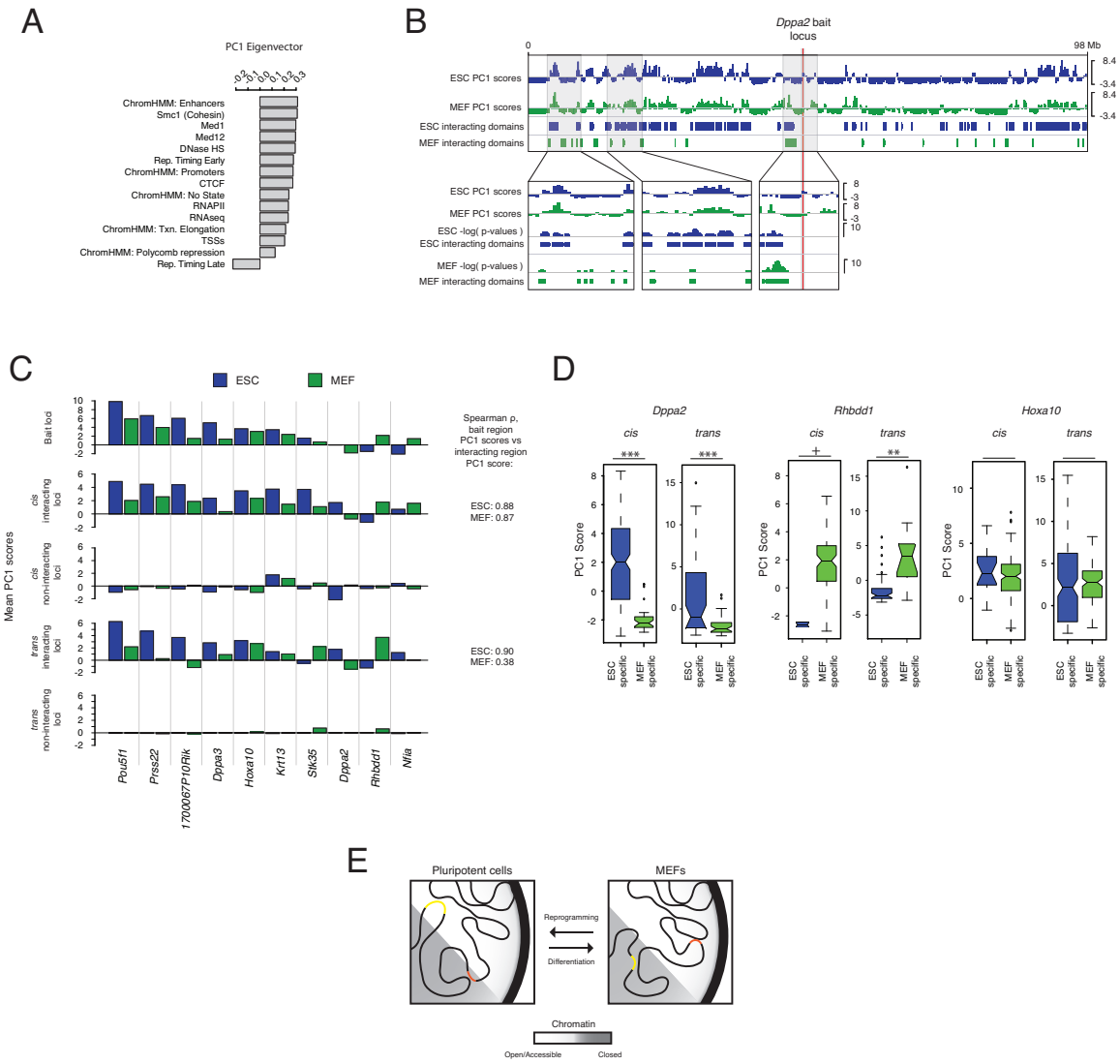

Figure 2.10 - Changes in open/closed chromatin character between ESCs and MEFs correspond to changes in interaction preferences

A, PCA was performed for concatenated ESC+MEF data, including the indicated features, resulting in the PC1 eigenvector depicted here.

B, Comparison of the intrachromosomal chromatin interactions of the *Dppa2* locus in ESCs and MEFs. Top to bottom: Integrative Genomics Viewer tracks showing the PC1 scores for chromosome 16 in ESCs (blue) and MEFs (green), and the interacting domains in ESCs and in MEFs determined by 4C-seq. Zoom-ins highlight the switch in PC1 bait character of the *Dppa2* locus upon differentiation (right) and the dramatic changes in interaction preferences (left two boxes).

C, Changes in the PC1 character of the bait region are associated with changes in the PC1 character of the interactome. Top to bottom: for each bait locus: Mean PC1 score for the 1Mb region centered on each listed bait's locus; mean PC1 score within interacting regions in *cis*; mean PC1 score within non-interacting regions in *cis*; mean PC1 score within interacting regions in *trans*; mean PC1 score within non-interacting regions in *trans*, for ESC data (blue) and MEF data (green). Spearman rho's give the correlation between the PC1 bait character and interactome character per cell type across all analyzed baits in *cis* and *trans*.

D, PC1 scores distributions of ESC and MEF-specific, significantly interacting domains of *Dppa2, Rhbdd1,* and *Hoxa10*. + p-value=0.099; **p-value <0.01; ***p-value <0.001; Wilcoxon rank-sum test.

E, Chromatin interaction model, wherein large-scale changes in chromatin interactions mirror changes in PC1 character upon ESC differentiation.

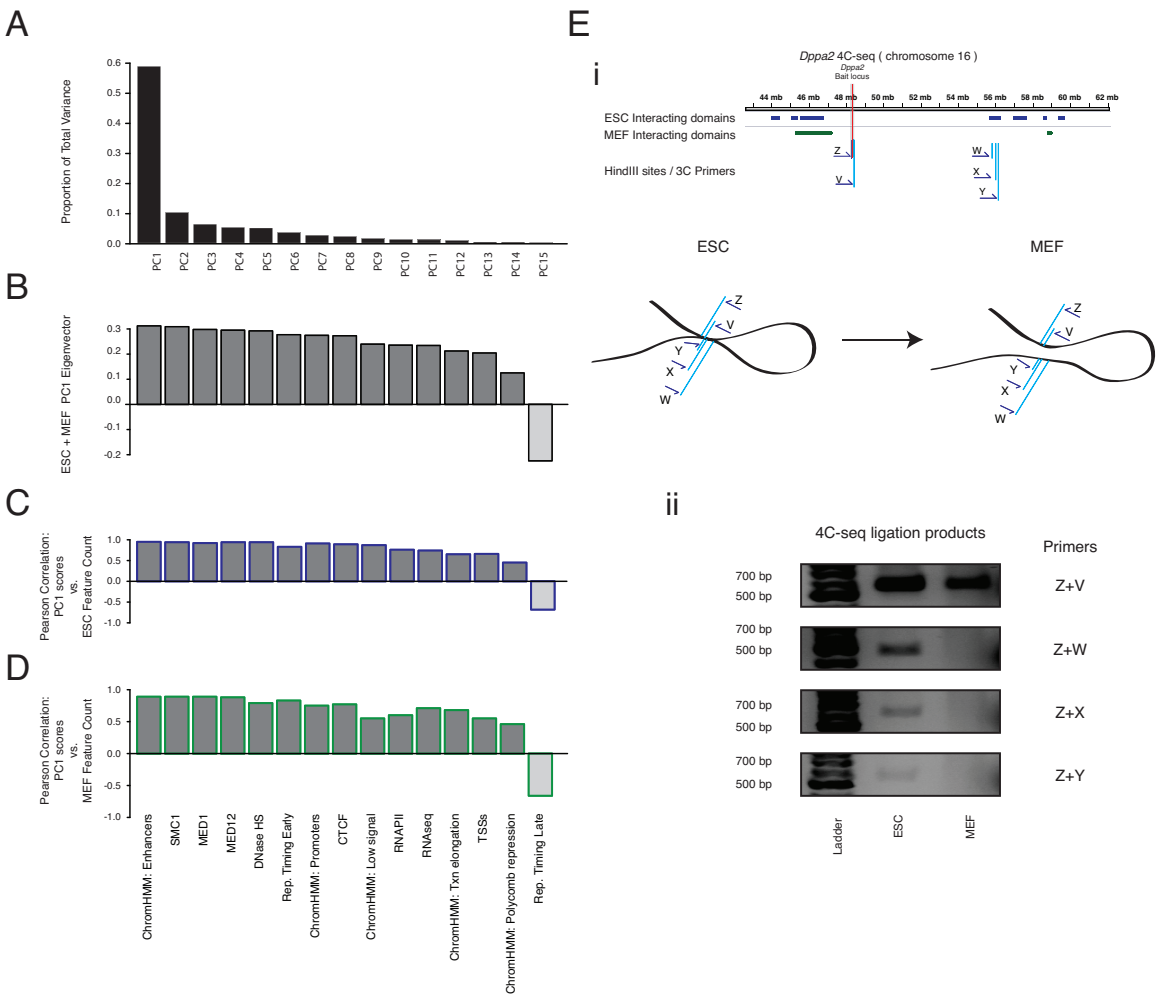

Figure 2.11 – Comparison of the spatial interactomes between ESCs and MEFs

A, PCA was performed on genomic features across ESCs and MEFs. Proportion of total variance in genomic features described by each principal component for the ESC+MEF PCA.

B, Genomic feature contribution to the ESC + MEF feature PC1 eigenvector.

C, Correlation of ESC genomic feature density with PC1 scores.

D, Correlation of MEF genomic feature density with PC1 scores.

E, (i) Schematic representation of the 3C experimental design to confirm the cell type-specificity of interactions with the *Dppa2* locus. At the top, a portion of chromosome 16 harboring the *Dppa2* locus is depicted, and a subset of interacting domains in ESCs and MEFs as defined by our 4C-seq analysis is indicated. Three ESC-specific interaction sites were examined (primers W-Y) by 3C in both MEFs and ESCs. A positive 3C control was designed to amplify a ligation product between the *Dppa2* locus (primer Z) and a proximal restriction site (primer V), which should be detectable in both ESCs and MEFs. (ii) 3C PCR results confirming the presence and

cell type-specificity of the ESC-specific ligation products Z/W, Z/X, and Z/Y. Note the presence of the positive control PCR product (Z+V) in both the ESC and MEF 3C experiments.

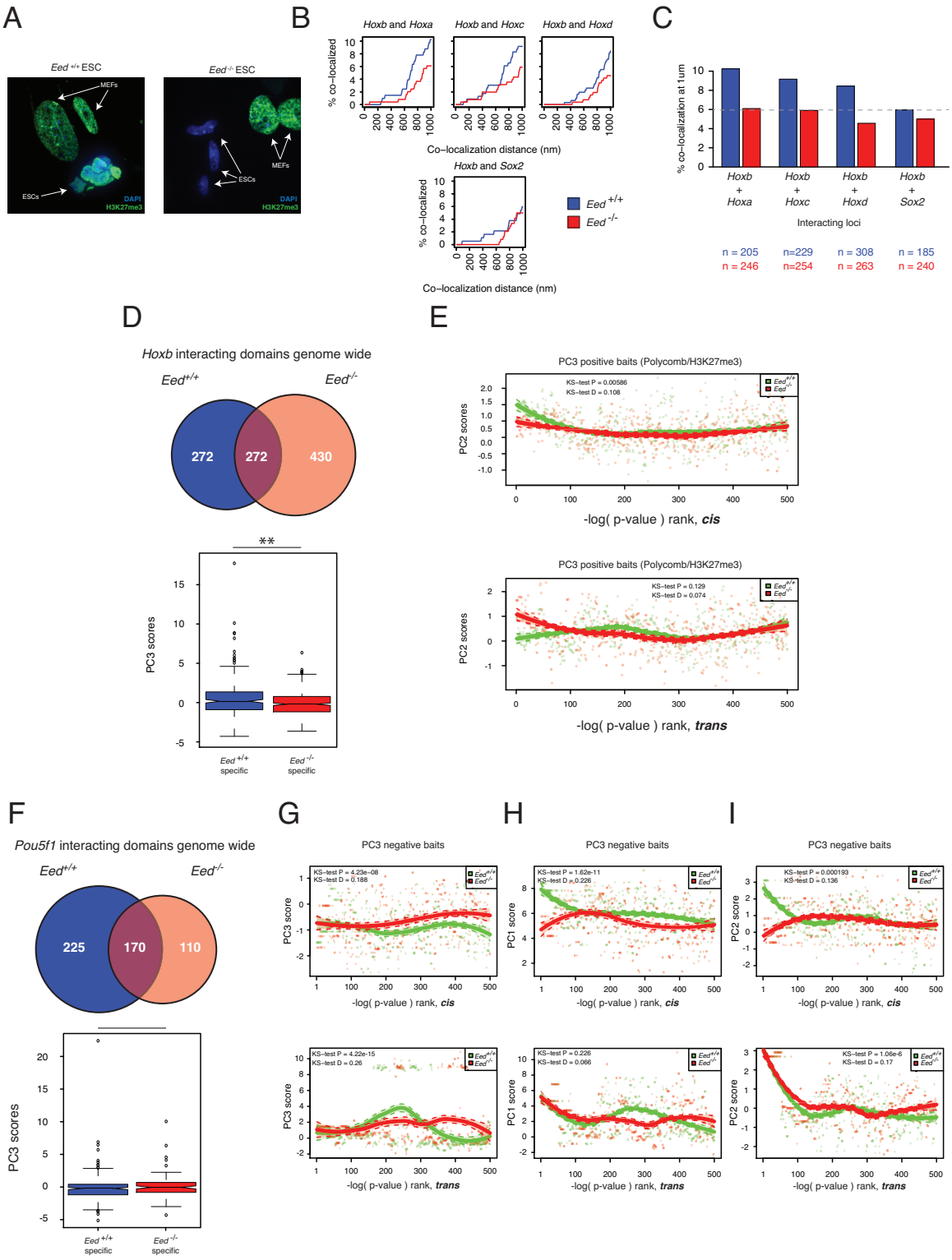

Figure 2.12 – Additional validation and characterization of interaction preferences in *Eed+/+* and *Eed-/-* ESCs

A, Immunostaining for H3K27me3 (green) in wildtype (*Eed*$^{+/+}$) and *Eed*$^{-/-}$ ESCs, grown on irradiated MEFs. DAPI staining in blue marks the nuclei. Note that Eed ablation leads to the loss of H3K27me3. Wildtype MEFs still stain positively for H3K27me3.

B. DNA FISH provides an independent confirmation of 4C-seq-defined differences in chromatin interactions between *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs**.** Cumulative distribution plots for interaction frequencies (y-axis) at different distances (x-axis) for the *trans* interactions measured between the *Hoxb* region and the individual *Hox* loci indicated in the figure (the sum of these individual plots is shown in Figure 2.13I-i), and, for comparison, between the *Hoxb* region and the *Sox2* region, which does not interact with *Hoxb* based on our 4C-seq data (same FISH analysis plot as shown in Figure 2.13I-i. The data for wildtype ESCs are shown in blue, and for mutant ESCs in red.

C, Co-localization frequencies at 1um for each pair of interactions listed in (B). The sum of these individual plots between *Hox* genes is shown in Figure 2.13-ii, and the *Hoxb-Sox2* analysis plot is the same as that shown in Figure 2.13I-ii for comparison. n = FISH signal pairs analyzed for each cell type.

D, Top: Overlap of interacting domains of the *Hoxb3* bait region between *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs genome-wide. Bottom: PC3 score distribution for the *Eed*$^{+/+}$- and the *Eed*$^{-/-}$ ESC-specific interacting domains. These data demonstrate that interacting domains that are specific for wildtype ESCs have significantly higher PC3 scores genome-wide than those that are specific for *Eed*$^{-/-}$ ESCs. ** p-value <= 0.01.

E, The *cis*- and *trans* interactomes of the six PC3-positive (Polycomb/H3K27me3 enriched) target bait loci *Hoxa10, Hoxb3, Hoxc4, Hox12, Pcdhb19, Tbx5* were ranked by -log(p-value) for both *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs, and, in each case, the 500 top ranked sites plotted against their average PC2 scores in wildtype ESCs, indicating minor changes in interaction preferences with regards to PC2 scores upon *Eed* ablation.

F, As in (D), except for the non-Polycomb, PC3-negative bait *Pou5f1*. *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESC specific interactions do not show significant differences in PC3 scores.

G-I, The *cis*- and *trans* interactomes of the two PC3-negative target bait loci *Pou5f1* and *Ptprg* were ranked by -log(p-value) for both *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs, and the 500 top ranked sites in *cis* (top) or *trans* (bottom) plotted against their average (G) PC3; (H) PC1; and (I) PC2 scores in wildtype ESCs. These data show that non-Polycomb targets do not show major changes in

interaction preferences with regards to the PC's. Kolmogorov-Smirnov test (KS) p-values shown with KS-test D value.

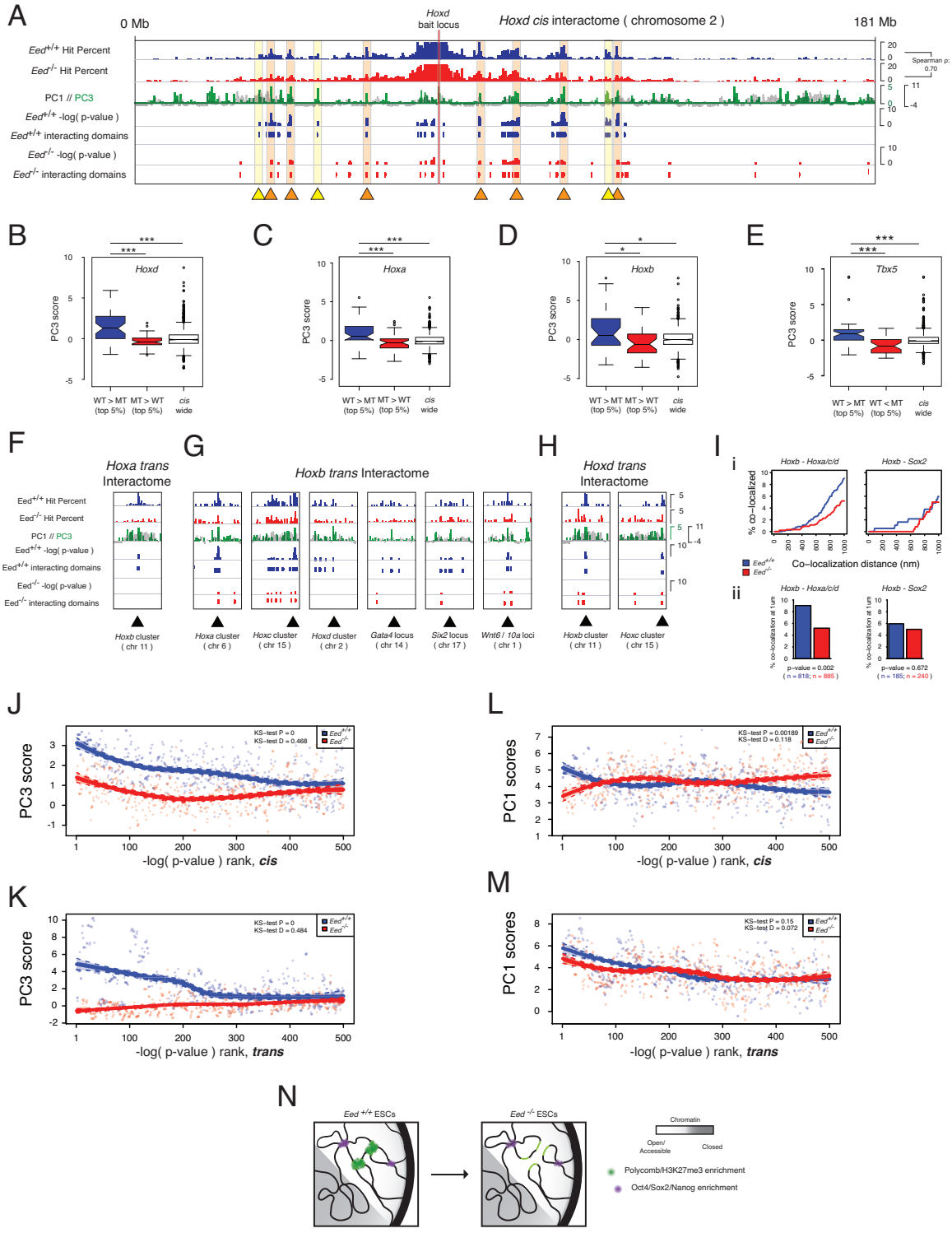

Figure 2.13 - *Eed* is required for the co-localization of PC3-positive genomic regions

A, Integrative Genomics Viewer tracks showing the *Hoxd12* interactome in *cis* in terms of hit percentage within 200kb windows, PC1 scores (black) overlaid with PC3 scores (green, positive values shown only), binomial test -log(p-values), and interacting domains for $Eed^{+/+}$ (blue) and

Eed$^{-/-}$ (red) ESCs. Regions that lose significant interactions with the *Hoxd12* bait region upon *Eed* ablation are marked with yellow triangles and shading; and those that do not lose interactions with the *Hoxd12* locus upon *Eed* ablation, but show a dramatic decrease in interaction strength are marked with orange triangles and shading. The Spearman *rho* compares the hit percentage in *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs.

B, *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs *Hoxd12* 4C-seq *cis*-interactome data shown in (A, hit percentage within 200kb windows) were subtracted and the 200kb windows within the top and bottom 5% of resulting values were used to define regions of the chromosome that showed stronger interactions in *Eed*$^{+/+}$ (WT>MT) or *Eed*$^{-/-}$ (MT>WT) ESCs, respectively. The PC3 score distribution of these genomic regions, as well as the *cis* chromosome-wide PC3 score distributions are shown. *** = p-value < 0.001, Wilcoxon rank-sum test.

C, As in (B), except for the *Hoxa10* 4C-seq interactome data in *cis*.

D, As in (B), except for the *Hoxb3* 4C-seq interactome data in *cis*, * = p-value < 0.05.

E, As in (B), except for the *Tbx5* 4C-seq interactome data in *cis*.

F-H, *Trans* interactions between the indicated (PC3-positive) *Hox* loci in *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs, displayed as in (A). *Hoxa* refers to the *Hoxa10* bait locus 4C-seq result, *Hoxb* to the *Hoxb3* locus 4C-seq result, *Hoxc* to the *Hoxc4* bait locus 4C-seq result, and *Hoxd* to the *Hoxd12* bait locus 4C-seq result.

I, DNA FISH analysis of *trans* interactions between *Hox* clusters. (i): Cumulative frequency distribution plots of co-localization frequencies (y-axis) between *Hoxb3* (chr11) and the other three *Hox* loci (*Hoxa10-chr6*, *Hoxc4-chr15*, *Hoxd12-chr2*) (left), and between *Hoxb3* and the *Sox2* (chr3) locus (right), with co-localization distances noted on the x-axis, measured in *Eed*$^{+/+}$ (blue) and *Eed*$^{-/-}$ (red) ESCs. Note that the background level co-localization frequencies between *Sox2* (which has a negative PC3 score (-0.16) and does not interact with *Hoxb3* based on our 4C-seq data) and *Hoxb3* does not change upon *Eed* ablation. (ii) Co-localization frequencies at 1um for *Hoxb3* and the other *Hox* loci (left) as well as for *Hoxb3* and *Sox2* (right) derived from (i). n = FISH signal pairs analyzed in both (i) and (ii); p-value = two-tailed Fisher exact test.

J, The *cis* interactomes of the six PC3-positive (Polycomb/H3K27me3 enriched) bait loci (*Hoxa10, Hoxb3, Hoxc4, Hoxd12, Pcdhb19, Tbx5*, see Figure 4G,ii) were ranked by -log(p-value) for both *Eed*$^{+/+}$ (blue) and *Eed*$^{-/-}$ (red) ESCs, and the 500 top sites plotted against their average PC3 scores in wildtype ESCs. Loess regression was used to generate a smooth curve,

and the Kolmogorov–Smirnov (KS) test to determine the probability that the two underlying one-dimensional probability distributions differ; D = KS-test D statistic).

K, As in (J), but in *trans*.

L, As in (J), except for PC1 scores.

M, As in (K), except for PC1 scores.

N, Chromatin interaction model, wherein in the absence of Eed, the frequency of interactions of regions with high PC3 scores is reduced, but large-scale chromosome conformation is largely conserved.

TABLES

Note: Tables 2.2, 2.3, and 2.4 are too large for inclusion in this document, and will be included a as a supplementary upload upon dissertation filing. Alternatively, they can be obtained by contacting Matthew Denholtz directly at matthew.denholtz@gmail.com.

	ESC	EED+/+ ESC	EED-/- ESC	iPSC	pre-iPSC	MEF
1700067P10Rik	✔			✔	✔	✔
Dppa2	✔			✔	✔	✔
Dppa3	✔			✔	✔	✔
Hoxa10	✔	✔	✔	✔	✔	✔
Hoxb		✔	✔			
Hoxc		✔	✔			
Hoxd		✔	✔			
Krt13	✔				✔	✔
Maoa	✔					
Nfia	✔			✔	✔	✔
Pcdhb19	✔	✔	✔		✔	
Pou5f1	✔	✔	✔	✔	✔	✔
Pou5f1_Control	✔					
Prss22	✔				✔	✔
Ptprg		✔	✔			
Rhbdd1	✔			✔	✔	✔
Stk35	✔			✔	✔	✔
Tbx5	✔	✔	✔	✔	✔	
Ugcg	✔			✔	✔	
Vegfc	✔					
Zfp42	✔					

Table 2.1 - Bait genes utilized by cell type. List of all bait genes whose interactomes were defined in this work and demarkations showing which baits were examined in each cell type

Bait gene	4C v Hi-C Sp
1700067P10Rik	0.73
Dppa2	0.62
Dppa3	0.88
Hoxa10	0.76
Krt131	0.87
Maoa	0.57
Nfia	0.69
Pcdhb19	0.67
Pou5f1	0.80
Prss22	0.88
Rhbdd1	0.71
Stk35	0.76
Tbx5	0.83
Ugcg	0.73
Vegfc	0.80
Zfp42	0.79
average	**0.76**
Pou5f1_Control	0.04

Table 2.5 - **4C-seq hit percent v. binned Hi-C read count** (Dixon et al.[42]). All pooled ESC 4C-seq data sets were correlated with their respective Hi-C defined interactomes. Spearman rank correlations across 200kb windows are reported.

Cell type	Feature	Citation
ESC	Eed	Tavares L, Dimitrova E, Oxley D, Webster J et al. RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. *Cell* 2012 Feb 17;148(4):664-78
ESC	Ezh2	Peng JC, Valouev A, Swigut T, Zhang J et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* 2009 Dec 24;139(7):1290-302
ESC	Suz12	Peng JC, Valouev A, Swigut T, Zhang J et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* 2009 Dec 24;139(7):1290-302
ESC	Max	Kim J, Woo AJ, Chu J, Snow JW et al. A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell* 2010 Oct 15;143(2):313-24
ESC	E2F4	Kim J, Woo AJ, Chu J, Snow JW et al. A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell* 2010 Oct 15;143(2):313-24
ESC	TBP	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	Smc1	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	Smc3	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	Med12	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	Med1	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	Ring1b	GEO accession number: GSM656523
ESC	CTCF	Handoko L, Xu H, Li G, Ngan CY et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011 Jun 19;43(7):630-8
ESC	PolI2	Kagey, et al. Nature. 2010 Sep 23;467(7314):430-5. Epub 2010 Aug 18
ESC	DNASEb	Encode consortium
ESC	Rep. Timing	Hiritani, Ryba, Itoh, Rathjen, Kulik, Papp, Fussner, Bazett-Jones, Plath, Dalton, Rthjen, Gilbert. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Research, 2010.
ESC	Nanog	This study
ESC	Oct4	This study
ESC	Sox2	This study
ESC	Klf4	This study
ESC	cMyc	This study
ESC	H3K4me1	This study
ESC	H3K4me2	This study
ESC	H3K4me3	This study
ESC	H3K27ac	This study
ESC	H3K27me3	This study
ESC	H3K36me3	This study
ESC	RNA-seq	This study
ESC	p300	This study
ESC	LaminB	Handoko L, Xu H, Li G, Ngan CY et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 2011 Jun 19;43(7):630-8
ESC	Mappability	Encode consortium
MEF	DNASEb	Encode consortium
MEF	PolII	GEO accesion number: GSM555138
MEF	CTCF	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010 Sep 23;467(7314):430-5
MEF	Med1	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010 Sep 23;467(7314):430-5
MEF	Med12	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010 Sep 23;467(7314):430-5
MEF	Smc1	Kagey MH, Newman JJ, Bilodeau S, Zhan Y et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010 Sep 23;467(7314):430-5
MEF	Rep. Timing	Hiritani, Ryba, Itoh, Rathjen, Kulik, Papp, Fussner, Bazett-Jones, Plath, Dalton, Rthjen, Gilbert. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Research, 2010.
MEF	H3K4me1	This study
MEF	H3K4me2	This study
MEF	H3K4me3	This study
MEF	H3K27ac	This study
MEF	H3K27me3	This study
MEF	H3K36me3	This study
MEF	RNA-seq	This study

Table 2.6 - **Linear genomic feature data sets.** Cell type, feature, and citation are listed. Those citations with the notation “This study” were generated by Kostas Chronis for the manuscript whose revision makes up Chapter 2.

Key:		
	full R2:	R-squared for log-log regression using all data points
	full B0:	Intercept for log-log regression using all data points
	full B1:	Slope for log-log regression using all data points
	proximal R2:	R-squared for log-log regression using 1-8 Mb data range
	proximal B0:	Intercept for log-log regression using 1-8 Mb data range
	proximal B1:	Slope for log-log regression using 1-8 Mb data range
	megabase R2:	R-squared for linear regression using 0-1 Mb data range
	megabase B0:	Intercept for regression for linear regression using 0-1 Mb data range
	megabase B1:	Slope for regression for linear regression using 0-1 Mb data range

Cell type	full R2	full B0	full B1	proximal R2	proximal B0	proximal B1	megabase R2	megabase B0	megabase B1
ESC baits	0.82	2.09	-0.97	0.99	1.75	-0.72	0.99	84.29	-30.01
ESC baits - Pou5f1	0.82	2.07	-0.96	0.99	1.76	-0.74	0.99	83.40	-29.53
ESC EED baits	0.78	1.87	-1.08	0.96	1.58	-0.86	1.00	89.95	-50.86
iPSC baits	0.81	1.80	-0.88	0.98	1.69	-0.94	0.99	84.94	-40.24
piPSC baits	0.78	1.59	-0.92	0.97	1.43	-0.96	0.99	76.26	-52.65
MEFs	0.79	1.43	-0.73	0.96	1.46	-1.01	0.98	82.09	-58.49

Table 2.7 - Binomial test empirical background model parameters. for each cell type listed. Key gives column headings.

REFERENCES

1 Splinter, E., de Wit, E., van de Werken, H. J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods*, doi:10.1016/j.ymeth.2012.04.009 (2012).

2 Zhang, Y. *et al*. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908-921, doi:10.1016/j.cell.2012.02.002 (2012).

3 Lieberman-Aiden, E. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

4 Sexton, T. *et al*. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010 (2012).

5 Denholtz, M. & Plath, K. Pluripotency in 3D: genome organization in pluripotent cells. *Current opinion in cell biology* **24**, 793-801, doi:10.1016/j.ceb.2012.11.001 (2012).

6 Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156 (1981).

7 Thomson, J. A. *et al*. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

8 Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940-954, doi:10.1016/j.cell.2011.01.032 (2011).

9 Rahl, P. B. *et al*. c-Myc regulates transcriptional pause release. *Cell* **141**, 432-445, doi:10.1016/j.cell.2010.03.030 (2010).

10 Boyer, L. A. *et al*. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).

11 Cao, R. *et al*. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039-1043, doi:10.1126/science.1076997 (2002).

12 Wang, H. *et al*. Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431**, 873-878, doi:10.1038/nature02985 (2004).

13 Eskeland, R. *et al*. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell* **38**, 452-464, doi:10.1016/j.molcel.2010.02.032 (2010).

14 Niwa, H., Ogawa, K., Shimosato, D. & Adachi, K. A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature* **460**, 118-122, doi:10.1038/nature08113 (2009).

15 Xu, R. H. *et al*. Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nature methods* **2**, 185-190, doi:10.1038/nmeth744 (2005).

16 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676, doi:10.1016/j.cell.2006.07.024 (2006).

17 Cremer, T. & Cremer, C. Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *European journal of histochemistry : EJH* **50**, 223-272 (2006).

18 Boyle, S. *et al*. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics* **10**, 211-219 (2001).

19 Warburton, D., Naylor, A. F. & Warburton, F. E. Spatial relations of human chromosomes identified by quinacrine fluorescence at metaphase. I. Mean interchromosomal distances and distances from the cell center. *Humangenetik* **18**, 297-306 (1973).

20 Parada, L. A., McQueen, P. G. & Misteli, T. Tissue-specific spatial organization of genomes. *Genome biology* **5**, R44, doi:10.1186/gb-2004-5-7-r44 (2004).

21 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews. Genetics* **2**, 292-301, doi:10.1038/35066075 (2001).

22 Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology* **2**, a003889, doi:10.1101/cshperspect.a003889 (2010).

23 Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & development* **18**, 1119-1130, doi:10.1101/gad.292104 (2004).

24 Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development* **20**, 2223-2237, doi:10.1101/gad.380906 (2006).

25 Andrulis, E. D., Neiman, A. M., Zappulla, D. C. & Sternglanz, R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature* **394**, 592-595, doi:10.1038/29100 (1998).

26 Kosak, S. T. *et al*. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science* **296**, 158-162, doi:10.1126/science.1068768 (2002).

27 Pickersgill, H. *et al*. Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature genetics* **38**, 1005-1014, doi:10.1038/ng1852 (2006).

28 Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243-247, doi:10.1038/nature06727 (2008).

29 Kumaran, R. I. & Spector, D. L. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of cell biology* **180**, 51-65, doi:10.1083/jcb.200706060 (2008).

30 Schirmer, E. C. & Foisner, R. Proteins that associate with lamins: many faces, many functions. *Experimental cell research* **313**, 2167-2179, doi:10.1016/j.yexcr.2007.03.012 (2007).

31 Mattout, A., Goldberg, M., Tzur, Y., Margalit, A. & Gruenbaum, Y. Specific and conserved sequences in D. melanogaster and C. elegans lamins and histone H2A mediate the attachment of lamins to chromosomes. *Journal of cell science* **120**, 77-85, doi:10.1242/jcs.03325 (2007).

32 Yao, J., Fetter, R. D., Hu, P., Betzig, E. & Tjian, R. Subnuclear segregation of genes and core promoter factors in myogenesis. *Genes & development* **25**, 569-580, doi:10.1101/gad.2021411 (2011).

33 Zullo, J. M. *et al*. DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell* **149**, 1474-1487, doi:10.1016/j.cell.2012.04.035 (2012).

34 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

35 Hakim, O. & Misteli, T. SnapShot: Chromosome confirmation capture. *Cell* **148**, 1068 e1061-1062, doi:10.1016/j.cell.2012.02.019 (2012).

36 Simonis, M. *et al*. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354, doi:10.1038/ng1896 (2006).

37 Zhao, Z. *et al*. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* **38**, 1341-1347, doi:10.1038/ng1891 (2006).

38 Dostie, J. *et al*. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309, doi:10.1101/gr.5571506 (2006).

39 Duan, Z. *et al*. A three-dimensional model of the yeast genome. *Nature* **465**, 363-367, doi:10.1038/nature08973 (2010).

40 Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* **30**, 90-98, doi:10.1038/nbt.2057 (2012).

41 Fullwood, M. J. *et al*. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).

42 Dixon, J. R. *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

43 Nora, E. P. *et al*. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

44 Handoko, L. *et al*. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics* **43**, 630-638, doi:10.1038/ng.857 (2011).

45 Shen, Y. *et al*. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

46 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

47 Wood, A. M. *et al*. Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. *Molecular cell* **44**, 29-38, doi:10.1016/j.molcel.2011.07.035 (2011).

48 Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).

49 Schmidt, D. *et al*. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335-348, doi:10.1016/j.cell.2011.11.058 (2012).

50 Noordermeer, D. *et al*. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol* **13**, 944-951, doi:10.1038/ncb2278 (2011).

51 Schoenfelder, S. *et al*. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics* **42**, 53-61, doi:10.1038/ng.496 (2010).

52 Bantignies, F. *et al*. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* **144**, 214-226, doi:10.1016/j.cell.2010.12.026 (2011).

53 Meshorer, E. *et al*. Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* **10**, 105-116, doi:10.1016/j.devcel.2005.10.017 (2006).

54 Efroni, S. *et al*. Global transcription in pluripotent embryonic stem cells. *Cell stem cell* **2**, 437-447, doi:10.1016/j.stem.2008.03.021 (2008).

55 Ahmed, K. *et al*. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PloS one* **5**, e10531, doi:10.1371/journal.pone.0010531 (2010).

56 Hiratani, I. *et al*. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome research* **20**, 155-169, doi:10.1101/gr.099796.109 (2010).

57 Hiratani, I. *et al*. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS biology* **6**, e245, doi:10.1371/journal.pbio.0060245 (2008).

58 Peric-Hupkes, D. *et al*. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell* **38**, 603-613, doi:10.1016/j.molcel.2010.03.016 (2010).

59 Minkovsky, A., Patel, S. & Plath, K. Concise review: Pluripotency and the transcriptional inactivation of the female Mammalian X chromosome. *Stem Cells* **30**, 48-54, doi:10.1002/stem.755 (2012).

60 Melcer, S. *et al*. Histone modifications and lamin A regulate chromatin protein dynamics in early embryonic stem cell differentiation. *Nature communications* **3**, 910, doi:10.1038/ncomms1915 (2012).

61 Kim, Y. *et al*. Mouse B-type lamins are required for proper organogenesis but not by embryonic stem cells. *Science* **334**, 1706-1710, doi:10.1126/science.1211222 (2011).

62 Capell, B. C. & Collins, F. S. Human laminopathies: nuclei gone genetically awry. *Nature reviews. Genetics* **7**, 940-952, doi:10.1038/nrg1906 (2006).

63 Tan-Wong, S. M., Wijayatilake, H. D. & Proudfoot, N. J. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. *Genes & development* **23**, 2610-2624, doi:10.1101/gad.1823209 (2009).

64 Capelson, M. *et al*. Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. *Cell* **140**, 372-383, doi:10.1016/j.cell.2009.12.054 (2010).

65 Kalverda, B., Pickersgill, H., Shloma, V. V. & Fornerod, M. Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell* **140**, 360-371, doi:10.1016/j.cell.2010.01.011 (2010).

66 Vaquerizas, J. M. *et al*. Nuclear pore proteins nup153 and megator define transcriptionally active regions in the Drosophila genome. *PLoS genetics* **6**, e1000846, doi:10.1371/journal.pgen.1000846 (2010).

67 Lupu, F., Alves, A., Anderson, K., Doye, V. & Lacy, E. Nuclear pore composition regulates neural stem/progenitor cell differentiation in the mouse embryo. *Dev Cell* **14**, 831-842, doi:10.1016/j.devcel.2008.03.011 (2008).

68 D'Angelo, M. A., Gomez-Cavazos, J. S., Mei, A., Lackner, D. H. & Hetzer, M. W. A change in nuclear pore complex composition regulates cell differentiation. *Developmental cell* **22**, 446-458, doi:10.1016/j.devcel.2011.11.021 (2012).

69 Bacher, C. P. *et al*. Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation. *Nat Cell Biol* **8**, 293-299, doi:10.1038/ncb1365 (2006).

70 Xu, N., Tsai, C. L. & Lee, J. T. Transient homologous chromosome pairing marks the onset of X inactivation. *Science* **311**, 1149-1152, doi:10.1126/science.1122984 (2006).

71 Masui, O. *et al*. Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation. *Cell* **145**, 447-458, doi:10.1016/j.cell.2011.03.032 (2011).

72 Zhang, L. F., Huynh, K. D. & Lee, J. T. Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. *Cell* **129**, 693-706, doi:10.1016/j.cell.2007.03.036 (2007).

73 Zhao, R., Bodnar, M. S. & Spector, D. L. Nuclear neighborhoods and gene expression. *Current opinion in genetics & development* **19**, 172-179, doi:10.1016/j.gde.2009.02.007 (2009).

74 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature genetics* **30**, 167-174, doi:10.1038/ng820 (2002).

75 Splinter, E. *et al*. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development* **25**, 1371-1383, doi:10.1101/gad.633311 (2011).

76 Agrelo, R. *et al*. SATB1 defines the developmental context for gene silencing by Xist in lymphoma and embryonic cells. *Dev Cell* **16**, 507-516, doi:10.1016/j.devcel.2009.03.006 (2009).

77 Cai, S., Han, H. J. & Kohwi-Shigematsu, T. Tissue-specific nuclear architecture and gene expression regulated by SATB1. *Nature genetics* **34**, 42-51, doi:10.1038/ng1146 (2003).

78 Cai, S., Lee, C. C. & Kohwi-Shigematsu, T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature genetics* **38**, 1278-1288, doi:10.1038/ng1913 (2006).

79 Yasui, D., Miyano, M., Cai, S., Varga-Weisz, P. & Kohwi-Shigematsu, T. SATB1 targets chromatin remodelling to regulate genes over long distances. *Nature* **419**, 641-645, doi:10.1038/nature01084 (2002).

80 Nechanitzsky, R., Dávila, A., Savarese, F., Fietze, S. & Grosschedl, R. Satb1 and Satb2 Are Dispensable for X Chromosome Inactivation in Mice. *Dev Cell* **23**, 866-871 (2012).

81 Wutz, A. & Agrelo, R. Response: The Diversity of Proteins Linking Xist to Gene Silencing. *Dev Cell* **23**, 680 (2012).

82 Monkhorst, K., Jonkers, I., Rentmeester, E., Grosveld, F. & Gribnau, J. X inactivation counting and choice is a stochastic process: evidence for involvement of an X-linked activator. *Cell* **132**, 410-421, doi:10.1016/j.cell.2007.12.036 (2008).

83 Maherali, N. *et al*. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell stem cell* **1**, 55-70, doi:10.1016/j.stem.2007.05.014 (2007).

84 Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nature reviews. Genetics* **9**, 129-140, doi:10.1038/nrg2295 (2008).

85 Haering, C. H., Farcas, A. M., Arumugam, P., Metson, J. & Nasmyth, K. The cohesin ring concatenates sister DNA molecules. *Nature* **454**, 297-301, doi:10.1038/nature07098 (2008).

86 Hadjur, S. *et al*. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460**, 410-413, doi:10.1038/nature08079 (2009).

87 Wendt, K. S. *et al*. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801, doi:10.1038/nature06634 (2008).

88 Parelho, V. *et al*. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433, doi:10.1016/j.cell.2008.01.011 (2008).

89 Rubio, E. D. *et al*. CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 8309-8314, doi:10.1073/pnas.0801273105 (2008).

90 Stedman, W. *et al*. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO journal* **27**, 654-666, doi:10.1038/emboj.2008.1 (2008).

91 Kagey, M. H. *et al*. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).

92 Schmidt, D. *et al*. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research* **20**, 578-588, doi:10.1101/gr.100479.109 (2010).

93 Faure, A. J. *et al*. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome research*, doi:10.1101/gr.136507.111 (2012).

94 Taatjes, D. J. The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends in biochemical sciences* **35**, 315-322, doi:10.1016/j.tibs.2010.02.004 (2010).

95 Levasseur, D. N., Wang, J., Dorschner, M. O., Stamatoyannopoulos, J. A. & Orkin, S. H. Oct4 dependence of chromatin structure within the extended Nanog locus in ES cells. *Genes & development* **22**, 575-580, doi:10.1101/gad.1606308 (2008).

96 Li, G. *et al*. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).

97 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26**, 11-24, doi:Doi 10.1101/Gad.179804.111 (2012).

98 Bickmore, W. A. & van Steensel, B. Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* **152**, 1270-1284, doi:Doi 10.1016/J.Cell.2013.02.001 (2013).

99 Gibcus, J. H. & Dekker, J. The Hierarchy of the 3D Genome. *Molecular cell* **49**, 773-782, doi:Doi 10.1016/J.Molce1.2013.02.011 (2013).

100 Nora, E. P., Dekker, J. & Heard, E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays : news and reviews in molecular, cellular and developmental biology*, doi:10.1002/bies.201300040 (2013).

101 Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Current opinion in cell biology* **25**, 387-394, doi:10.1016/j.ceb.2013.02.005 (2013).

102 Wei, Z. *et al*. Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell stem cell* **13**, 36-47, doi:10.1016/j.stem.2013.05.010 (2013).

103 Osborne, C. S. *et al*. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics* **36**, 1065-1071, doi:10.1038/ng1423 (2004).

104 Papantonis, A. *et al*. TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *The EMBO journal* **31**, 4404-4414, doi:10.1038/emboj.2012.288 (2012).

105 Hakim, O. *et al*. Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome research* **21**, 697-706, doi:10.1101/gr.111153.110 (2011).

106 Hakim, O. *et al*. Spatial congregation of STAT binding directs selective nuclear architecture during T-cell functional differentiation. *Genome research* **23**, 462-472, doi:10.1101/gr.147652.112 (2013).

107 Bernstein, B. E. *et al*. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326, doi:10.1016/j.cell.2006.02.041 (2006).

108 Mikkelsen, T. S. *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

109 Ernst, J. *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

110 Chen, X. *et al*. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:10.1016/j.cell.2008.04.043 (2008).

111 Sridharan, R. *et al*. Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136**, 364-377, doi:10.1016/j.cell.2009.01.001 (2009).

112 Boyer, L. A. *et al*. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956, doi:10.1016/j.cell.2005.08.020 (2005).

113 Loh, Y. H. *et al*. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature genetics* **38**, 431-440, doi:10.1038/ng1760 (2006).

114 Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049-1061, doi:10.1016/j.cell.2008.02.039 (2008).

115 Miele, A., Gheldof, N., Tabuchi, T. M., Dostie, J. & Dekker, J. Mapping chromatin interactions by chromosome conformation capture. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 21**, Unit 21 11, doi:10.1002/0471142727.mb2111s74 (2006).

116 Nichols, J. *et al*. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**, 379-391 (1998).

117 Feldman, N. *et al*. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat Cell Biol* **8**, 188-194, doi:10.1038/ncb1353 (2006).

118 Takebayashi, S., Dileep, V., Ryba, T., Dennis, J. H. & Gilbert, D. M. Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 12574-12579, doi:10.1073/pnas.1207185109 (2012).

119 Apostolou, E. *et al*. Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell stem cell* **12**, 699-712, doi:10.1016/j.stem.2013.04.013 (2013).

120 Koche, R. P. *et al*. Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell stem cell* **8**, 96-105, doi:10.1016/j.stem.2010.12.001 (2011).

121 Polo, J. M. *et al*. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617-1632, doi:10.1016/j.cell.2012.11.039 (2012).

122 Kim, J. *et al*. A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. *Cell* **143**, 313-324, doi:10.1016/j.cell.2010.09.010 (2010).

123 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).

124 Imakaev, M. *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).

125 Whyte, W. A. *et al*. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

126 Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F. & Orlando, V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat Cell Biol* **9**, 1167-1174, doi:10.1038/ncb1637 (2007).

127 Tolhuis, B. *et al*. Interactions among Polycomb domains are guided by chromosome architecture. *PLoS genetics* **7**, e1001343, doi:10.1371/journal.pgen.1001343 (2011).

128 Montgomery, N. D. *et al*. The murine polycomb group protein Eed is required for global histone H3 lysine-27 methylation. *Current biology : CB* **15**, 942-947, doi:10.1016/j.cub.2005.04.051 (2005).

129 Chamberlain, S. J., Yee, D. & Magnuson, T. Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. *Stem Cells* **26**, 1496-1505, doi:10.1634/stemcells.2008-0102 (2008).

130 Phillips-Cremins, J. E. *et al*. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).

131 Rinn, J. L. *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323, doi:10.1016/j.cell.2007.05.022 (2007).

132 Engreitz, J. M. *et al*. The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science*, doi:10.1126/science.1237973 (2013).

133 Tanizawa, H. *et al*. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research* **38**, 8164-8177, doi:10.1093/nar/gkq955 (2010).

134 Ye, Q., Callebaut, I., Pezhman, A., Courvalin, J. C. & Worman, H. J. Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *The Journal of biological chemistry* **272**, 14983-14989 (1997).

135 Solovei, I. *et al*. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584-598, doi:10.1016/j.cell.2013.01.009 (2013).

136 Muthurajan, U. M., McBryant, S. J., Lu, X., Hansen, J. C. & Luger, K. The linker region of macroH2A promotes self-association of nucleosomal arrays. *The Journal of biological chemistry* **286**, 23852-23864, doi:10.1074/jbc.M111.244871 (2011).

137 Cowieson, N. P., Partridge, J. F., Allshire, R. C. & McLaughlin, P. J. Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis. *Current biology : CB* **10**, 517-525 (2000).

138 Csink, A. K. & Henikoff, S. Genetic modification of heterochromatic association and nuclear organization in Drosophila. *Nature* **381**, 529-531, doi:10.1038/381529a0 (1996).

139 Moissiard, G. *et al*. MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* **336**, 1448-1451, doi:10.1126/science.1221472 (2012).

140 Gaspar-Maia, A. *et al*. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-868, doi:10.1038/nature08212 (2009).

141 Fraser, P., Pruzina, S., Antoniou, M. & Grosveld, F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes & development* **7**, 106-113 (1993).

142 Cantor, A. B. & Orkin, S. H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**, 3368-3376, doi:10.1038/sj.onc.1205326 (2002).

143 Lin, Y. C. *et al*. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology* **13**, 1196-1204, doi:10.1038/ni.2432 (2012).

144 de Wit, E. *et al*. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **0**, 5 (2013).

145 Chess, A., Simon, I., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**, 823-834 (1994).

146 Clowney, E. J. *et al*. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell* **151**, 724-737, doi:10.1016/j.cell.2012.09.043 (2012).

147 Magklara, A. *et al*. An epigenetic signature for monoallelic olfactory receptor expression. *Cell* **145**, 555-570, doi:10.1016/j.cell.2011.03.040 (2011).

148 Lettice, L. A., Hill, A. E., Devenney, P. S. & Hill, R. E. Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Human molecular genetics* **17**, 978-985, doi:10.1093/hmg/ddm370 (2008).

149 Amano, T. *et al*. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**, 47-57, doi:10.1016/j.devcel.2008.11.011 (2009).

150 Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* **12**, 283-293, doi:10.1038/nrg2957 (2011).

151 Duncan, I. W. Transvection effects in Drosophila. *Annual review of genetics* **36**, 521-556, doi:10.1146/annurev.genet.36.060402.100441 (2002).

152 Muller, M., Hagstrom, K., Gyurkovics, H., Pirrotta, V. & Schedl, P. The mcp element from the Drosophila melanogaster bithorax complex mediates long-distance regulatory interactions. *Genetics* **153**, 1333-1356 (1999).

153 Guttman, M. *et al*. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300, doi:10.1038/nature10398 (2011).

154 Chan, C. S., Rastelli, L. & Pirrotta, V. A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. *The EMBO journal* **13**, 2553-2564 (1994).

155 Sing, A. *et al*. A vertebrate Polycomb response element governs segmentation of the posterior hindbrain. *Cell* **138**, 885-897, doi:10.1016/j.cell.2009.08.020 (2009).

156 Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140**, 99-110, doi:10.1016/j.cell.2009.12.022 (2010).

157 Bantignies, F., Grimaud, C., Lavrov, S., Gabut, M. & Cavalli, G. Inheritance of Polycomb-dependent chromosomal interactions in Drosophila. *Genes & development* **17**, 2406-2420, doi:10.1101/gad.269503 (2003).

158 Grimaud, C. *et al*. RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124**, 957-971, doi:10.1016/j.cell.2006.01.036 (2006).

159 Cheutin, T. & Cavalli, G. Progressive polycomb assembly on H3K27me3 compartments generates polycomb bodies with developmentally regulated motion. *PLoS genetics* **8**, e1002465, doi:10.1371/journal.pgen.1002465 (2012).

160 Ferraiuolo, M. A. *et al*. The three-dimensional architecture of Hox cluster silencing. *Nucleic acids research* **38**, 7472-7484, doi:10.1093/nar/gkq644 (2010).

161 Tiwari, V. K. *et al*. PcG proteins, DNA methylation, and gene repression by chromatin looping. *PLoS biology* **6**, 2911-2927, doi:10.1371/journal.pbio.0060306 (2008).

162 Lee, T. I. *et al*. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

163 Sridharan, R. *et al*. Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1gamma in reprogramming to pluripotency. *Nat Cell Biol* **15**, 872-882, doi:10.1038/ncb2768 (2013).

164 Morin-Kensicki, E. M., Faust, C., LaMantia, C. & Magnuson, T. Cell and tissue requirements for the gene eed during mouse gastrulation and organogenesis. *Genesis* **31**, 142-146 (2001).

165 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

166 van de Werken, H. J. *et al*. 4C technology: protocols and data analysis. *Methods in enzymology* **513**, 89-112, doi:10.1016/B978-0-12-391938-0.00004-5 (2012).

167 R: A language and environment for statistical computing v. 2.13.0 (R Foundation for Statistical Computing, Vienna, Austria., 2011).

168 Parkhomchuk, D. *et al*. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research* **37**, e123, doi:10.1093/nar/gkp596 (2009).

169 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

170 Wagschal, A., Delaval, K., Pannetier, M., Arnaud, P. & Feil, R. Chromatin Immunoprecipitation (ChIP) on Unfixed Chromatin from Cells and Tissues to Analyze Histone Modifications. *CSH protocols* **2007**, pdb prot4767, doi:10.1101/pdb.prot4767 (2007).

171 Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

172 Gue, M., Messaoudi, C., Sun, J. S. & Boudier, T. Smart 3D-FISH: automation of distance analysis in nuclei of interphase cells by image processing. *Cytometry A* **67**, 18-26 (2005).

173 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671–675 (2012).