Transcriptional and Structural Control of Cell Identity Genes

By

Zi Peng Fan

B.S. Biochemistry
Brandeis University, 2007

Submitted to the Program in Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy in Computational and Systems Biology
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015

Signature of Author Signature redacted
Program in Computational and Systems Biology
May 22nd, 2015

Certified by Signature redacted
Richard A. Young
Professor of Biology
Thesis Supervisor

Accepted by Signature redacted
Christopher B. Burge
Chairman, Department Computational and Systems Biology

Transcriptional and Structural Control of Cell Identity Genes

By
Zi Peng Fan

Submitted to the Program in Computational and Systems Biology
on May 22nd, 2015, in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy in Computational and Systems Biology

ABSTRACT

Mammals contain a wide array of cell types with distinct functions, yet nearly all cell types have the same genomic DNA. How the genetic instructions in DNA are selectively interpreted by cells to specify various cellular functions is a fundamental question in biology. This thesis work describes two genome-wide studies designed to study how transcriptional control of gene expression programs defines cell identity.

Recent studies suggest that a small number of transcription factors, called "master" transcription factors, dominate the control of gene expression programs. These master transcription factors and the transcriptional regulatory circuitry they produce, however, are not known for all cell types. Ectopic expression of these factors can, in principle, direct transdifferentiation of readily available cells into medically relevant cell types for applications in regenerative medicine. Limited knowledge of these factors is a roadblock to generation of many medically relevant cell types. Chapter 2 presents a study in which a novel computational approach was undertaken to generate an atlas of candidate master transcriptional factors for 100+ human tissue/cell types. The candidate master transcription factors in retinal pigment epithelial (RPE) cells were then used to guide the investigation of the regulatory circuitry of RPE cells and to reprogram human fibroblasts into functional RPE-like cells.

Master transcription factors define cell-type-specific gene expression through binding to enhancer elements in the genome. These enhancer-bound transcription factors regulate genes by contacting target gene promoters via the formation of DNA loops. It is becoming increasingly clear that transcription factors operate and regulate gene expression within a larger three-dimensional (3D) chromatin architecture, but these structures and their functions are poorly understood. Chapter 3 presents a study in which Cohesin ChIA-PET data was generated to identify the local chromosomal structures at both active and repressed genes across the genome in embryonic stem cells. The results led to the discovery of functional insulated neighborhood structures that are formed by two CTCF interaction sites occupied by Cohesin. The integrity of these looped structures contributes to the transcriptional control of super-enhancer-driven active genes and repressed genes encoding lineage-specifying developmental regulators.

Thesis supervisor: Richard A. Young
Title: Professor of Biology

Acknowledgements

I am very fortunate to have the opportunity to spend a wonderful 6-year period at MIT. The experience has made me a better researcher and scientist. Thank you Chris Burge, the CSB program, my classmates, and especially program administrators Bonnie Whang and Jacqueline Carota. I want to take this opportunity to wish all the best to my friends and colleagues at MIT and in Boston.

Thank you my advisor Rick Young for your support and guidance for my research projects and for pushing me to think harder and to think bigger. Thank you to my colleague and mentor Tony Lee for giving me advice and sharing his wisdom on issues both inside and outside of the lab.

I would like to thank my committee members Prof. Laurie Boyer and Chris Burge for advice and guidance for my research projects. I would like to thank Prof. Len Zon for giving me the opportunities to collaborate on a number of interesting projects. I would also like to thank Prof. Zhiping Weng for serving on my thesis defense committee.

I am very lucky to have the opportunities to work with many very talented scientists, Jill Dowen, Denes Hnisz, Alla Sigova, Ana D'Alessio, Abe Weintraub, Lars Anders, and Xiong Ji in the lab. I have learned a lot from them. I would also like to thank Lee Lawton and Charles Lin for teaching me about bioinformatics when joining the lab. Thank you to David Orlando, Charles Lin, Garrett Frampton, Brian Abraham, and BaRC for building world-class bioinformatics infrastructure in the lab. Lee Lawton, Jessica Reddy, Daniel Dadon, Abe Weintraub, Evan Cohick, and Jurian Schuijers have made the lab a fun place to work. Finally, all the members of the Young lab have contributed to my scientific and personal growth in different ways and your hard work inspires me.

A special thank you to my fiancée Siyang Su for her love and support. I spent the toughest and happiest time together with her during the time in graduate school. She brought me the best food and desserts I can ever ask for.

Most importantly, I'd like to thank my parents Xiao Hua Huo and Fudong Fan for their sacrifice, support, confidence, and love. They always encourage me to take on challenges and have the utmost confidence in me. I love you.

Table of Contents

Chapter 1

Introduction

Preface

Gene regulation is the process by which the genetic instructions stored in the DNA are selectively processed and interpreted by the cells. Understanding the regulation of gene expression is one of the fundamental goals of biological research. In my thesis work, I have developed and used computational methods to interrogate large-scale genome-wide datasets in order to predict key regulators of cell-type-specific gene expression and to study the relationship between chromosome structure and gene regulation. In the first chapter of my thesis I will provide a brief overview about transcriptional regulation by RNA polymerase II and three-dimensional chromosome structure. I first introduce cis-regulatory elements and components of transcription apparatus. I next discuss and highlight some insights into how a small number of transcription factors dominate the control of cell-type-specific gene expression programs. I then describe different levels of organization of the chromosome structures. In the second chapter, I describe a computational approach to generate an atlas of candidate master transcriptional regulators for a broad spectrum of human cells. The candidate regulators of retinal pigment epithelial (RPE) cells are used to guide the investigation of the transcriptional regulatory circuitry of RPE cells and

to reprogram human fibroblasts into RPE-like cells. In the third chapter, I describe a computational pipeline to analyze and visualize the sequencing results from genome-wide chromatin interaction data and its use to produce a genome-wide map of Cohesin-associated DNA loops that include enhancer-promoter loops as well as larger loop structures. This map reveals that super-enhancer-driven genes and polycomb-repressed genes frequently occur in "insulated neighborhoods". These neighborhoods are formed by large DNA loops that are co-bound by Cohesin and CTCF. Perturbation experiments suggest these neighborhoods serve to maintain proper expression of genes within and outside of the loop. In the final chapter, I present some unanswered questions and discuss some possible approaches to address these questions.

Transcriptional regulation: an overview

The regulation of gene expression is fundamental to cell-type-specific cellular function. In a typical human or mouse cell type, roughly 60%-70% of protein-coding genes are transcribed (Ramskold et al. 2009, Lee and Young 2013). This set of actively transcribed genes, often called the gene expression program, is transcribed by RNA polymerase II and largely defines cell identity. The control of gene expression programs involves specific DNA sequences and the regulatory proteins and RNA species that interact with them. This control also involves structural features of chromosomes.

Gene control depends largely on regulatory information encoded in four types of regulatory sequences: core promoter elements that contain the

transcription start site (Smale and Kadonaga 2003), promoter-proximal elements (Lenhard, Sandelin, and Carninci 2012), enhancer elements (Bulger and Groudine 2011), and insulator elements (West, Gaszner, and Felsenfeld 2002). Transcription factors regulate gene expression by binding to specific sequences in promoter proximal and enhancer elements, recruiting chromatin regulators to help generate an appropriate local chromatin state, and recruiting the transcription apparatus to the core promoter (reviewed in (Levine 2010, Bulger and Groudine 2011, Ong and Corces 2011, Zaret and Carroll 2011, Spitz and Furlong 2012, Lee and Young 2013, Slattery et al. 2014, Heinz et al. 2015)). Enhancer-bound transcription factors are thought to regulate their target genes by forming DNA loops in order to come into close proximity to the promoter of a target gene. Insulator elements are thought to block these DNA loop interactions between specific enhancer elements and potential target genes (Geyer and Corces 1992, Cai and Levine 1995, West, Gaszner, and Felsenfeld 2002). The mechanisms that allow insulators to block enhancer-promoter interactions are not well understood, but have been postulated to involve the ubiquitously-expressed DNA binding factor CTCF.

The ~2 meters of genomic DNA in mammalian cells is packaged into a nucleus of less than 10nm, and there are multiple levels of structural organization within chromosomes that allow this to occur (Misteli 2007, Gibcus and Dekker 2013). At the smallest level of organization, approximately 140bp of DNA is tightly wrapped around a nucleosome consisting of two molecules each of histone proteins H2A, H2B, H3, and H4 (Kornberg and Lorch 1999). At the next

level of organization, sites within this nucleosomal DNA, also called a 10nm fiber, form loops. Some of these loops originate from the interaction of proteins associated with enhancer and promoter elements (Sanyal et al. 2012). These enhancer-promoter DNA loop interactions are generally confined within regions of the genome called topologically associated domains (TADs) (Dixon et al. 2012, Nora et al. 2012), which are the local portions of the genome, averaging 0.8Mb, that tend to be in close contact; these TADs tend to be shared by most cell types. In interphase chromosomes, the TADS are organized into 2 types of megabase-scale compartments, termed A and B (Lieberman-Aiden et al. 2009). A compartments are "open", gene-rich, generally transcriptionally active; B compartments are "closed", gene-poor, and generally transcriptionally silent. Some relationships between gene regulation and chromosome structure are just beginning to be understood, and are addressed in more detail below.

Regulatory elements in the genome

Core Promoter Elements

Promoters are sequences flanking the transcription start site (TSS) of a gene and are generally defined to be the sequences that direct the initiation of transcription. The canonical core sequence elements can include the TATA box, B recognition element (BRE), initiator (Inr) element, motif ten element (MTE), and the downstream promoter element (DPE) (Juven-Gershon et al. 2008, Roy and Singer 2015). Promoters are often found to contain one or more different core promoter elements in different combinations. Different combinations of core

promoter elements likely reflect differential usage of these regulatory sequences in transcriptional control and the diversity of the assembly of transcription machinery (Decker and Hinton 2013). These promoter elements are bound by components of the general transcription apparatus, which include general transcription factors (GTFs) and RNA polymerase II (Roeder 1996, Lee and Young 2000).

Promoter-Proximal Elements

Some promoter-proximal elements often overlap with core promoter elements, but we will describe them here as elements that are located within several hundred bp of the TSS and that are bound by transcription factors or the paused transcription apparatus. A number of TFs have been noted that tend to bind in promoter-proximal regions, including c-MYC and SP1 (Dynan and Tjian 1983, Rahl et al. 2010). These factors may contribute to the recruitment or stability of the general transcription apparatus at the promoters, and c-MYC is thought to participate in RNA polymerase II pause release (Rahl et al. 2010).

Some promoter-proximal sequence elements may also contribute to transcriptional control via promoter-proximal RNA polymerase II pausing (Hendrix et al. 2008). Genome-wide studies suggested that promoter-proximal pausing is widespread (Zeitlinger 2007, Core, Waterfall, and Lis 2008, Rahl et al. 2010). In metazoans, the majority of protein-coding genes show evidence of transcription initiation, but for some of these, there is no evidence of elongation (Muse 2007, Guenther et al. 2007) and it has emerged that RNA polymerase II generally

pauses after synthesis of 20-60 bases near promoters (Adelman and Lis 2012). Promoter-proximal pausing is now thought to be an important regulatory step in RNAPII transcription which among other things, facilitates rapid and synchronous transcriptional responses upon exposure to transcriptional activation signals (Zeitlinger 2007, Muse 2007, Core, Waterfall, and Lis 2008, Rahl et al. 2010, Gilchrist 2010, Adelman and Lis 2012).

Distal enhancer elements

Enhancers are DNA elements that are distal to gene promoters and have the potential to enhance the basal transcription levels of target genes. The first enhancer element described was a DNA element from Simian virus 40 (SV40), which was shown to increase the expression of T-antigen and a b-globin reporter gene (Banerji, Rusconi, and Schaffner 1981). Enhancers were subsequently found in in many metazoan genomes and are now thought to be the primary determinant of tissue-specific gene expression (reviewed in (Spitz and Furlong 2012, Buecker and Wysocka 2012, Lee and Young 2013, Heinz et al. 2015)). Enhancers can be located at hundreds of bases to mega-bases from promoters (Banerji, Rusconi, and Schaffner 1981, Lettice et al. 2003). Enhancers are thought to regulate their target genes by coming into close proximity of the promoter of their target gene by forming DNA loop interactions (Tolhuis et al. 2002, Vakoc et al. 2005, Fullwood et al. 2009, Sanyal et al. 2012, Arnold et al. 2013). Therefore, the selective usage of enhancers, and subsequent regulation of specific target genes, is a critical component of the control of cell identity.

Distal enhancers serve as binding sites for a broad array of sequence-specific transcription factors encoded in the genome (reviewed in (Spitz and Furlong 2012, Buecker and Wysocka 2012, Lee and Young 2013, Heinz et al. 2015)). Multiple transcription factors are generally bound at any one enhancer (Chen et al. 2008, Kim et al. 2008, Yan et al. 2013, Cheng et al. 2014), and the combinatorial binding properties provide several useful gene control functions. Cooperative interactions between multiple transcription factors, each of which binds a small portion of the enhancer DNA sequence, permits synergistic and combinatorial effects that differ at different enhancers (Maniatis et al. 1998, Carey 1998, Segal et al. 2008). Combinatorial binding of cell-type specific transcription factors can also allow a single transcription factor to participate in multiple cell-type specific gene expression programs. The transcription factor Oct4, for example, can occupy distinct sets of enhancers in two closely related cell types – a embryonic stem cells and epiblast stem cells - depending on the expression level of its binding partners (Factor et al. 2014, Buecker et al. 2014). Furthermore, some transcription factors – especially those that are involved in transmitting signals from developmental signaling pathways – take advantage of cooperative interactions with other TFs in order to regulate the appropriate genes. Signaling-dependent transcription factors tend to bind enhancers occupied by lineage-specific transcription factors (Trompouki et al. 2011, Mullen et al. 2011).

Transcription factor binding at enhancers leads to the recruitment of transcriptional co-factors and in many cases, the recruitment of RNA polymerase II (RNAP II) and transcription at enhancers (Kim et al. 2010). The transcriptional

cofactors are defined as factors that play general roles in gene control but do not have their own DNA-binding capability, and include the Mediator/Cohesin complex (Kagey et al. 2010), histone acetyl-transferases (HAT) such as p300 and CREB-binding protein (CBP) (Wang et al. 2009), and chromatin remodelers such as the transcription activator BRG1 complex and the SWI/SNF complexes (Euskirchen et al. 2011, Morris et al. 2014). A specific chromatin signature characterized by DNase I hypersensitivity and specific covalent modifications (methylation and acetylation) of histone tails can be found at enhancers (Rivera and Ren 2013). This chromatin signature is produced by TF binding and recruitment of specific cofactors and is frequently used to identify putative enhancer elements.

Insulator elements

Insulators are DNA elements that have the ability to insulate a gene from regulatory influences (West, Gaszner, and Felsenfeld 2002). Insulator elements were first discovered in *Drosophila* when the DNA elements *scs* and *scs'* were found to mark the chromatin boundaries of a heat shock gene. Two insulator-binding proteins zeste-white (Zw5) and boundary element associated factor (BEAF) were subsequently discovered (Zhao, Hart, and Laemmli 1995, Gaszner, Vazquez, and Schedl 1999). Two regulatory functions were proposed for insulators based on the genetic studies in *Drosophila*. In some cases, insulators insulate a gene from aberrant activation by blocking the DNA loop interactions between enhancer elements and gene promoters (Kellum and Schedl 1991, Geyer and Corces 1992). In other cases, insulators insulate a gene from aberrant

repression by acting as act as a barrier at the boundaries between transcriptionally active and transcriptionally repressive chromatin (Sun and Elgin 1999).

CTCF is the only known insulator protein encoded in the mammalian genome (Bell, West, and Felsenfeld 1999) and is highly conserved in higher eukaryotes (Ohlsson, Renkawitz, and Lobanenkov 2001). It is an 11-zinc finger protein that binds to a core consensus DNA sequence CCCTC. CTCF was first discovered and isolated on the basis of its binding within the promoter-proximal regulatory regions of avian *c-myc* gene (Lobanenkov et al. 1990, Klenova et al. 1993). Mouse and human CTCF was subsequently discovered to bind at conserved regions of c-myc genes (Filippova et al. 1996). In these studies, CTCF was described as transcriptional repressor because of its ability to repress gene expression in reporter assays. It was subsequently shown that CTCF confers diverse regulatory functions in a context-dependent manner (reviewed in (Phillips and Corces 2009, Merkenschlager and Odom 2013)), including enhancer blocking, transcriptional activation/repression, insulation, imprinting, X chromosome inactivation, and formation of chromatin domain structures. It is now thought that CTCF confers many, if not most, of these functions by creating DNA loops (Phillips and Corces 2009). These DNA loops can confer insulator functions by creating topological structures that constrain the interactions between regulatory elements and genes.

Components of the transcription apparatus

The control of transcription initiation and elongation is carried out largely by transcription factors, transcriptional co-factors, and RNA polymerase II together with a set of general transcription factors (reviewed by (Roeder 1996, Lee and Young 2000, Kornberg 2007). In addition, the transition from transcription initiation to processive elongation is thought to involve RNA polymerase II pausing and pause release, and there are several regulators that contribute to this process (Adelman and Lis 2012).

Transcription factors

A role for trans-acting factors in gene control was first proposed in models that emerged from pioneering genetic studies of the lac operon in bacteria in the 1960s (Jacob and Monod 1961). Genes encoded in the lac operon are required for the metabolism of lactose and the model that emerged from the studies of Jacob and Monod can be described as follows. In the absence of lactose, lac repressor, a transcription factor, binds to an operator sequence at the lac gene promoter to inhibit transcription. In the presence of lactose and the absence of glucose, lac repressor is inhibited and is dissociated from the operator sequence at the promoter. This leads to transcriptional activation of lac operon by transcription-activating catabolite activator protein (CAP). These observations demonstrate a fundamental concept of gene control in which transcription factors bind to DNA sequence elements and recruit protein complexes that activate or

repress the transcription of a gene. This concept has provided a foundation for understanding gene control in all organisms.

Transcription factors recognize specific DNA sequences through contacts with both DNA bases as well as the three-dimensional structure of DNA (reviews in (Rohs et al. 2010, Slattery et al. 2014)). Transcription factors typically bind to 6-12 bp DNA sequences. The preferred sequences recognized by the transcription factor DNA binding domains are known as "DNA motifs". Transcription factors generally bind to DNA sequences based on the chemical complementarity between the major and/or minor grooves of the DNA double helix and the amino acid side chains on the surfaces of the transcription factors ((Badis et al. 2009, Stormo and Zhao 2010, Jolma et al. 2013)). This form of transcription factor-DNA recognition is known as "base readout". In addition, interactions between transcription factor and DNA depend on the three-dimensional structures of both macromolecules. Transcription factors can recognize the structural information of the DNA double helix, such as DNA shape (Joshi et al. 2007, Rohs et al. 2009, Gordan et al. 2013), bending(Stella, Cascio, and Johnson 2010) and unwinding (Chen et al. 2013). This form of transcription factor-DNA recognition is known as "shape readout".

Transcription factors control gene expression mainly at the steps of transcription initiation and elongation (reviewed in (Spitz and Furlong 2012, Lee and Young 2013)). Most transcription factors are thought to contribute to transcription initiation by recruiting co-activators, which in turn bind the general transcription apparatus at core promoters, thus forming DNA loops. Some

transcription factors, including, c-MYC (Eberhardy and Farnham 2002, Rahl et al. 2010) and NF-kb (Barboric et al. 2001) bind to core promoters and recruit positive elongation factor b (P-TEFb), whose activity is necessary for efficient pause release. The functional contributions of transcription factor binding to transcription initiation and elongation are not always readily distinguishable. For example, many transcription factors interact with various subunits of the Mediator complex, which has been implicated in the control of both initiation and elongation (reviewed in (Taatjes 2010, Yin and Wang 2014)). It is therefore possible that these transcription factors also contribute to the control of both transcription initiation and transcription elongation in a context-specific fashion.

Transcriptional cofactors: the Mediator complex

The Mediator complex, also known as Mediator, is required for full transcriptional activity of gene expression *in vitro* and *in vivo* (reviewed in (Kornberg 2005, Lee and Young 2000, Roeder 2005, Malik and Roeder 2005, Conaway and Conaway 2011, Allen and Taatjes 2015)). Mediator is a large multi-subunit protein complex that acts as a central scaffold that interacts with and bridges DNA-binding transcription factors, general transcription factors and RNAPII. This cofactor is made of more than 20 core subunits, and individual subunits are targeted by specific transcription factors and are linked to specific transcriptional responses (reviewed in (Taatjes 2010, Yin and Wang 2014)).

During transcription initiation, Mediator promotes the assembly of an enhancer-promoter complex by binding both enhancer-bound transcription

factors and promoter-bound pre-initiation complex and by interacting with the cohesin-loading protein NIPBL, which loads cohesin, which in turn contributes to stability of the looped complexes (Kagey et al. 2010). During transcription elongation, Mediator helps recruit multiple components of super-elongation complex (SEC) to stimulate RNAPII elongation (Takahashi et al. 2011, Ebmeier and Taatjes 2010). Thus the Mediator complex is thought to be a centralized "hub" for transcriptional regulation.

Transcriptional Cofactors: P300-CBP Coactivator Family

Many transcription factors have been shown to interact with p300 and CBP, which have similar structures and functions and are thus considered to be within the same family (reviewed in (Chan and La Thangue 2001, Shikama, Lyon, and LaThangue 1997)). p300 and CBP contain multiple well-defined protein interaction domains, including the nuclear receptor interaction domain, the CREB and MYB interaction domain, the interferon response binding domain, cysteine/histidine regions, a histone acetyltransferase domain, a bromodomain that binds acetylated lysines and a PHD finger motif. When transcription factors recruit P300/CBP, these coactivators produce the histone H3K27Ac modification that is used widely as a marker for active enhancers and is among a variety of histone acetylation events that are thought to contribute to "open" chromatin (Heintzman et al. 2009, Creyghton et al. 2010, Rada-Iglesias et al. 2011).

RNA polymerase II (RNAPII)

Eukaryotic core RNA polymerase II (RNAPII) is a highly conserved, multi-protein enzymatic complex made of 10-12 subunits (reviewed in (Myer and Young 1998, Lee and Young 2000, Kornberg 2007, Grunberg and Hahn 2013, Sainsbury, Bernecky, and Cramer 2015)). There are three different types of RNA polymerase responsible for transcription of eukaryotic genomes (Vannini and Cramer 2012). RNAPII mainly transcribes protein-coding genes as well as non-coding cis-regulatory sequences; whereas RNAPI transcribes the ribosomal RNA genes, and RNAPIII transcribes genes encoding tRNAs and other non-coding RNAs

One important regulatory feature of RNAPII is the highly conserved carboxy-terminal repeat domain (CTD) of the largest RNAPII subunit, which plays important roles in various stages of transcription and in coupling transcription to pre-mRNA processing (reviewed in (Buratowski 2003, Hsin and Manley 2012)). The CTD of vertebrate RNAPII contains 52 tandem heptad repeats of the amino acid sequence YSPTSPS (Chapman et al. 2008). The phosphorylation state at serine 2 and serine 5 of these tandem repeats is tightly linked to the transcription stage of RNA polymerase II. These tandem repeats are phosphorylated at serine 5 by the CDK7 subunit of the general transcription factor TFII-H and at serine 2 by the CDK9 subunit of the positive transcription elongation factor (P-TEFb) (discussed below). RNA polymerase II with hypo-phosphorylated CTD preferentially associates with the transcription pre-initiation complex (PIC) (Lu et al. 1991). After PIC assembly, the hypo-phosphorylated CTD of RNAPII is

phosphorylated at Serine 5 by the CDK7 subunit of TFIIH (Lu et al. 1992). Phosphorylation of Serine 2 at RNAPII CTD by P-TEFb occurs during the transition from initiation to elongation (Marshall and Price 1992, Marshall et al. 1996), leading to the recruitment of enzymes responsible for pre-mRNA 5' capping (McCracken et al. 1997). Pre-mRNA 5' capping may be required for RNAPII transitions from initiation to elongation (Moore and Proudfoot 2009).

General Transcription Factors

Transcription of protein-coding genes by RNAPII involves three main stages: transcription initiation, transcription elongation, and transcription termination (reviewed in (Roeder 1996, Lee and Young 2000, Kornberg 2007)). During transcription initiation, RNAPII assembles at the core promoter with general transcription factors (GTFs; also known as Basal Transcription factors) to form a pre-initiation complex (PIC). The general transcription factors, which include TFII-B, TFII-D, TFII-E, TFII-F, and TFII-H, are essential for RNAPII binding to promoters and allow low levels of transcription at core promoters *in vitro* (also known as "basal transcription").

RNA Polymerase II Pause and Pause-Release Factors

Following transcription initiation, RNAPII generally pauses after synthesis of 20-60 bases near promoters (Adelman and Lis 2012). Two proteins complexes play key roles in the promoter-proximal pausing of RNAPII by interacting directly with RNAPII complex (Wu et al. 2003): the negative elongation factor complex (NELF) and the DRB sensitivity-inducing factor (DSIF) (Wada et al. 1998,

Yamaguchi et al. 1999). The release of paused RNAPII requires the recruitment of the positive transcription elongation factor, P-TEFb. The P-TEFb is a cyclin dependent kinase comprised of Cyclin T and CDK9 (Marshall and Price 1995). P-TEFb phosphorylate NELF, DSIF and the Ser2 residue of the RNAPII CTD heptad repeat, resulting in the transition of RNAPII into the processive elongation mode (Wada et al. 1998, Peterlin and Price 2006). After pause release, the processive RNAPII continues transcription elongation across the gene body and terminates shortly after transcription of signals for cleavage and polyadenylation at the end of the gene.

Transcriptional control of cell identity

Cell-type-specific gene expression programs are defined by active transcription of genes required for specialized cellular functions and repression of genes that specifies other lineages. The key specificity determinants of gene expression programs are transcription factors. In a typical cell type, hundreds of transcription factors are expressed. However, studies of transcriptional control of gene expression programs suggest that a small number of key transcription factors, called "master" transcription factors, dominate the control of gene expression programs (Graf and Enver 2009, Orkin and Hochedlinger 2011, Young 2011, Lee and Young 2013).

Genetic and cellular reprogramming studies demonstrate that a small number of transcription factors are required for both the establishment and maintenance of cell-type-specific gene expression programs. Genetic

experiments have shown that the loss of specific transcription factors can cause loss of cell identity and can stimulate lineage-switching or differentiation into another cell type. In embryonic stem cells, the loss of expression of master transcription factors Oct4 and Sox2 results in differentiation and thus loss of the pluripotent cell state (Chambers and Smith 2004, Masui et al. 2007, Wang et al. 2012). In mature B-cells, genetic ablation of master transcription factor Pax5 results in cell dedifferentiation to an early progenitor state and aberrant expression of genes from the T-cell lineages (Cobaleda, Jochum, and Busslinger 2007). In some cases, the loss of these factors can lead to apoptosis or other forms of cell death. For example, transcription factor Gata1, which is essential in red blood cells, is shown to regulate genes important for red blood cell functions and also to suppress apoptosis (Weiss and Orkin 1995).

Cellular reprograming experiments have shown that ectopic expression of a small set of transcription factors has the ability to reprogram cell identity. Weintraub and colleagues first showed that ectopic expression of the basic helix-loop-helix (bHLH) transcription factor MyoD is sufficient to convert fibroblasts into contracting myocytes (Lassar, Paterson, and Weintraub 1986, Davis, Weintraub, and Lassar 1987). More recently, Yamanaka and colleagues showed that ectopic expression of four transcription factors, Oct4, Sox2, c-Myc, and Klf4 could reprogram somatic cells into induced pluripotent stem cells that were similar to embryonic stem cells (Takahashi and Yamanaka 2006). Similar types of reprogramming studies have led to the identification of various transcription

factors capable of inducing new cell states for nearly a dozen cell types (Lee and Young 2013).

Studies of transcriptional control of embryonic stem cells (ESCs) have provided insights into how a small set of master transcription factors control cell-type-specific gene expression programs (Young 2011, Lee and Young 2013). First, the genes encoding the master transcription factors Oct4, Sox2, and Nanog are expressed at high levels and their expression tends to be cell-type restricted. Second, Oct4, Sox2, and Nanog occupy a substantial fraction of active enhancers and recruit multiple transcriptional co-factors to their target genes (Chen et al. 2008, Marson et al. 2008). Third, master transcription factors frequently form positive interconnected auto-regulatory loops that have been termed the "core regulatory circuitry" of ESCs (Boyer et al. 2005). These core regulatory circuitry auto-regulatory loops have been observed in many additional well-studied cell types, including hepatocytes (Odom et al. 2006), T cell acute lymphoblastic leukemia cells (Sanda et al. 2012), hematopoeitc stem cells and erythroid cells (Novershtern et al. 2011). This type of network structure depicts transcription factors regulating the expression of their own genes as well as those of the other master transcription factors. Such network structures have been shown to reinforce and increase the stability of gene expression programs (Alon 2007), and also likely explain why gene expression programs can be maintained throughout the cell cycle.

Chromosome Structure

There are multiple levels of structural organization within chromosomes that allow ~2 meters of genomic DNA in mammalian cells to be packaged into a nucleus of less than 10nm (Gibcus and Dekker 2013, Gorkin, Leung, and Ren 2014). These levels of structure include nucleosomes, DNA loops that connect enhancers and promoters or create domains called insulated neighborhoods (Dowen et al. 2014), and then larger regions called topologically associated domains (TADs). Each of these is discussed in more detail below.

Nucleosomes

Nucleosomes consist of approximately 147bp of DNA wrapped around a histone octamer containing two molecules of each of the histone proteins H2A, H2B, H3 and H4 (Kornberg 1974, Oudet, Grossbellard, and Chambon 1975). Each of these core histone proteins is composed of a highly structured C-terminal histone domain that binds tightly to DNA and an unstructured N-terminal tail that protrudes from the nucleosome core. The N-terminal tails are enriched for lysine and arginine residues, which can be subjected to a wide array of post-translational modifications, including acetylation, methylation, phosphorylation, and many others (reviewed in (Kornberg and Lorch 1999, Campos and Reinberg 2009, Kouzarides 2007)).

Nucleosome occupancy or modification plays various roles in gene regulation. Nucleosome occupancy of a transcription factor binding site can reduce the ability of some transcription factors to bind the site. In contrast,

transcription factor-occupied enhancers tend to have limited nucleosome occupancy, which is due largely to the action of ATP-dependent chromosome remodeling complexes that are recruited by some transcription factors; these remodeling complexes can use the energy of ATP hydrolysis to mobilize nucleosomes (Hargreaves and Crabtree 2011). Histone modifications can alter the interactions among nucleosomes to render chromatin more "open" to transcription factor binding, or to produce binding sites that are recognized by transcriptional coactivators or co-repressors (reviewed in (Kornberg and Lorch 1999, Kouzarides 2007, Campos and Reinberg 2009)).

Specific histone modifications occur in nucleosomes that occupy active cis-regulatory elements and their associated genes. Nucleosomes with histone H3K27ac and H4K4me1 modifications are found at active enhancer elements (Creyghton et al. 2010, Rada-Iglesias et al. 2011). Histone H3K4me3, H3K79me2, and H3K36me3 modifications are found within transcriptionally active genes. Histone H3K4me3 modification occurs in nucleosomes immediately downstream of promoters of genes that experience initiation by RNAPII (Bernstein et al. 2002, Pokholok et al. 2005, Guenther et al. 2008). Histone H3K79me2 and K3K36me3 modification occurs within the bodies of genes that are transcribed by elongating RNAPII; H3K79me2 modification occurs in nucleosomes near the promoter regions of genes (Feng et al. 2002), and H3K36me3 occurs in nucleosomes that are further downstream of transcribed genes (Sun et al. 2005, Bannister et al. 2005, Pokholok et al. 2005).

There are other histone modifications that are associated with gene repression. Nucleosomes with histone H3K9me3 and H3K27me3 modifications occupy repressed genes: H3K9me3 tends to occur at transcriptionally silent genes or in repetitive DNA elements (Lachner et al. 2001), whereas H3K27me3 occurs in genes that, in embryonic stem cells, encode lineage-specific developmental regulators that are repressed in ES cells but poised for rapid activation during differentiation (Boyer et al. 2006, Lee et al. 2006, Orkin and Hochedlinger 2011, Young 2011).

DNA loop interactions

Among the DNA loops that have been described, two types of DNA loops play important roles in gene regulation and they are discussed here. One involves DNA loops that connect enhancers and the promoters and the other involves loops that fully encompass one or more genes with their regulatory elements and that act to constrain those elements to act within the DNA loop. These latter types of loops are called insulated neighborhoods (Dowen et al. 2014).

Transcription factors bind enhancers, recruit coactivators such as Mediator, which in turn binds RNA polymerase II at promoters, thus forming a DNA loop between enhancers and promoters. The Cohesin loading factor Nipbl co-localizes with Mediator, providing a means to load Cohesin and thus contribute to the stability of DNA loops between enhancers and promoters (Kagey et al. 2010). In some cases, the enhancer-promoter loops may also

involve interaction between CTCF bound at the enhancer or promoter, or both sites (Majumder et al. 2008, Liu et al. 2011, Handoko et al. 2011, Seitan, Krangel, and Merkenschlager 2012).

Large DNA loop interactions involving two CTCF-bound sites also occur at many sites that are not enhancers or promoters, and these can insulate a gene from an enhancer or encompass one or more genes with their enhancers (Phillips and Corces 2009, Handoko et al. 2011, Dowen et al. 2014). CTCF forms homodimers and other multimers *in vitro* (Moon et al. 2005), which explains how DNA loops can be formed between two CTCF bound regions. CTCF also physically interacts with Cohesin through the C-terminal region of CTCF and the SA2 subunit of Cohesin (Xiao, Wallace, and Felsenfeld 2011). One of the best studied examples of insulation from an enhancer involves the imprinted *Igf2/H19* locus. On the maternal allele, a DNA loop interaction between two CTCF sites is formed to block the *Igf2* promoter from accessing a downstream enhancer (Kurukuti et al. 2006).

We have recently reported that large DNA loop interactions involving two CTCF bound sites can encompass a super-enhancer and its target gene and create an insulated domain (Dowen et al. 2014). Loss of either of the CTCF sites leads to altered expression of the normal super-enhancer driven gene and the super-enhancer will then activate genes that are normally located outside of the CTCF-bounded loop.

Topologically associating domains

One important structural feature of chromosome organization is the self-interacting topologically associated domains (TADs) (Dixon et al. 2012, Nora et al. 2012). These domains are hundreds of kilobases in size. They tend to be shared by most cell types and also tend to be conserved across species (Dixon et al. 2012). Chromatin interaction maps generated by 5C and HiC techniques suggest that DNA loop interactions tend to be confined within these TADs (Dixon et al. 2012, Nora et al. 2012). The boundaries of TADs are regions across which relatively few DNA-DNA interactions occur. In addition, the boundaries are typically enriched for both CTCF and Cohesin (Dixon et al. 2012, Sofueva et al. 2013).

TADs may contribute to gene control by constraining interactions between regulatory elements and genes (Gibcus and Dekker 2013, Gorkin, Leung, and Ren 2014). The conservation of TADs across cell types implies that most cell-type-specific DNA loop interactions (e.g. enhancer-promoter DNA loops) should occur at the sub-TAD level (Phillips-Cremins et al. 2013). Several lines of evidence support the model that TAD boundaries tend to be shared by most cell types, whereas sub-TAD structure varies by cell type. First, studies of the transcriptional control of the mouse HoxD cluster suggest that enhancer-associated interactions are confined by TADs such that the enhancers can only interact with a subset of HoxD genes (Andrey et al. 2013). Second, the expression of genes within TADs is more correlated than genes between TADs during development (Nora et al. 2012). Third, genetic deletion of TAD boundary

regions can lead to inappropriate DNA interactions and de-regulation of gene expression within TADs (Nora et al. 2012, Zuin et al. 2014). These results suggest that TADs represent a level of chromosome organization that is connected to regulation of gene expression.

Concluding Remarks

In those cell types where the control of gene expression is relatively well understood, a small number of key transcription factors, called "master" transcription factors, are known to dominate the control of the gene expression program. These master transcription factors are known for only a small fraction of all human cell types, and it would be valuable to identify candidate master transcription factors for all cell types. Indeed, an atlas of master transcription factors could guide exploration of the core transcriptional regulatory circuitry of clinically important cell types, and could also facilitate advances in direct reprogramming for these cell types. In chapter 2, I present a study in which a novel computational approach was undertaken to generate an atlas of candidate master transcriptional factors for 100+ human tissue/cell types. The candidate master transcription factors in retinal pigment epithelial (RPE) cells were then used to guide the investigation of the regulatory circuitry of RPE cells and to reprogram human fibroblasts into functional RPE-like cells.

Recent studies indicate that the genome is organized into topologically associated domains (TADs), which contribute to gene control by constraining interactions between regulatory elements and genes. Knowledge that super-

enhancers drive expression of genes with prominent roles in cell identity led us to investigate the sub-TAD structure associated with these unusual elements. In chapter 3, I present a study in which Cohesin ChIA-PET data was generated to identify local chromosomal structures at both active and repressed genes in embryonic stem cells. The results led to the discovery of functional insulated neighborhood structures that are formed by two CTCF interaction sites occupied by Cohesin. The integrity of these looped structures contributes to the transcriptional control of super-enhancer-driven active genes and repressed genes encoding lineage-specifying developmental regulators. This study demonstrates that sub-TAD structures formed by CTCF-CTCF interactions can contribute the transcriptional control of cell identity genes.

Acknowledgements

I wish to thank members of the Young lab, especially Rick Young, Tony Lee, and Jessica Reddy, Jill Dowen, and Jurian Schuijers for helpful comments during the preparation of this chapter.

References

Adelman, K., and J. T. Lis. 2012. "Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans." *Nature Rev. Genet.* 13:720-731.

Allen, B. L., and D. J. Taatjes. 2015. "The Mediator complex: a central integrator of transcription." *Nature Reviews Molecular Cell Biology* 16 (3):155-166. doi: DOI 10.1038/nrm3951.

Alon, U. 2007. "Network motifs: theory and experimental approaches." *Nature Reviews Genetics* 8 (6):450-461. doi: Doi 10.1038/Nrg2102.

Andrey, G., T. Montavon, B. Mascrez, F. Gonzalez, D. Noordermeer, M. Leleu, D. Trono, F. Spitz, and D. Duboule. 2013. "A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs." *Science* 340 (6137):1195-+. doi: ARTN 1234167 DOI 10.1126/science.1234167.

Arnold, C. D., D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath, and A. Stark. 2013. "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq." *Science* 339 (6123):1074-1077. doi: Doi 10.1126/Science.1232542.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Y. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. 2009. "Diversity and Complexity in DNA Recognition by Transcription Factors." *Science* 324 (5935):1720-1723. doi: Doi 10.1126/Science.1162327.

Banerji, J., S. Rusconi, and W. Schaffner. 1981. "Expression of a [beta]-globin gene is enhanced by remote SV40 DNA sequences." *Cell* 27:299-308.

Bannister, A. J., R. Schneider, F. A. Myers, A. W. Thorne, C. Crane-Robinson, and T. Kouzarides. 2005. "Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes." *Journal of Biological Chemistry* 280 (18):17732-17736. doi: DOI 10.1074/jbc.M500796200.

Barboric, M., R. M. Nissen, S. Kanazawa, N. Jabrane-Ferrat, and B. M. Peterlin. 2001. "NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II." *Mol Cell* 8 (2):327-37.

Bell, A. C., A. G. West, and G. Felsenfeld. 1999. "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." *Cell* 98 (3):387-396. doi: Doi 10.1016/S0092-8674(00)81967-4.

Bernstein, B. E., E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides, and S. L. Schreiber. 2002. "Methylation of histone H3 Lys 4 in coding regions of active genes." *Proceedings of the National Academy of Sciences of the United States of America* 99 (13):8695-8700.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. 2005. "Core transcriptional regulatory circuitry in human embryonic stem cells." *Cell* 122 (6):947-56. doi: S0092-8674(05)00825-1 [pii] 10.1016/j.cell.2005.08.020.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch. 2006. "Polycomb complexes repress developmental regulators in murine embryonic stem cells." *Nature* 441 (7091):349-53. doi: nature04733 [pii] 10.1038/nature04733.

Buecker, C., R. Srinivasan, Z. X. Wu, E. Calo, D. Acampora, T. Faial, A. Simeone, M. J. Tan, T. Swigut, and J. Wysocka. 2014. "Reorganization of Enhancer Patterns in Transition from Naive to Primed Pluripotency." *Cell Stem Cell* 14 (6):838-853. doi: Doi 10.1016/J.Stem.2014.04.003.

Buecker, C., and J. Wysocka. 2012. "Enhancers as information integration hubs in development: lessons from genomics." *Trends Genet* 28 (6):276-84. doi: 10.1016/j.tig.2012.02.008.

Bulger, M., and M. Groudine. 2011. "Functional and mechanistic diversity of distal transcription enhancers." *Cell* 144:327-339.

Buratowski, S. 2003. "The CTD code." *Nature Structural Biology* 10 (9):679-680. doi: Doi 10.1038/Nsb0903-679.

Cai, H., and V. Levine. 1995. "Modulation of Enhancer-Promoter Interactions by Insulators in the Drosophila Embryo." *Nature* 376 (6540):533-536. doi: DOI 10.1038/376533a0.

Campos, E. I., and D. Reinberg. 2009. "Histones: annotating chromatin." *Annu Rev Genet* 43:559-99. doi: 10.1146/annurev.genet.032608.103928.

Carey, M. 1998. "The enhanceosome and transcriptional synergy." *Cell* 92 (1):5-8. doi: Doi 10.1016/S0092-8674(00)80893-4.

Chambers, I., and A. Smith. 2004. "Self-renewal of teratocarcinoma and embryonic stem cells." *Oncogene* 23 (43):7150-7160. doi: Doi 10.1038/Sj.Onc.1207930.

Chan, H. M., and N. B. La Thangue. 2001. "p300/CBP proteins: HATs for transcriptional bridges and scaffolds." *Journal of Cell Science* 114 (13):2363-2373.

Chapman, R. D., M. Heidemann, C. Hintermair, and D. Eick. 2008. "Molecular evolution of the RNA polymerase II CTD." *Trends Genet* 24 (6):289-96. doi: 10.1016/j.tig.2008.03.010.

Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. 2008. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell* 133 (6):1106-17. doi: S0092-8674(08)00617-X [pii] 10.1016/j.cell.2008.04.043.

Chen, Y. H., X. J. Zhang, A. C. D. Machado, Y. Ding, Z. C. Chen, P. Z. Qin, R. Rohs, and L. Chen. 2013. "Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion." *Nucleic Acids Research* 41 (17):8368-8376. doi: Doi 10.1093/Nar/Gkt584.

Cheng, Y., Z. H. Ma, B. H. Kim, W. S. Wu, P. Cayting, A. P. Boyle, V. Sundaram, X. Y. Xing, N. Dogan, J. J. Li, G. Euskirchen, S. Lin, Y. Lin, A. Visel, T. Kawli, X. Q. Yang, D. Patacsil, C. A. Keller, B. Giardine, A. Kundaje, T. Wang, L. A. Pennacchio, Z. P. Weng, R. C. Hardison, M. P. Snyder, and Mouse ENCODE Consortium. 2014. "Principles of regulatory information conservation between mouse and human." *Nature* 515 (7527):371-+. doi: Doi 10.1038/Nature13985.

Cobaleda, C., W. Jochum, and M. Busslinger. 2007. "Conversion of mature B cells into T cells by dedifferentiation to uncommitted progenitors." *Nature* 449 (7161):473-U8. doi: Doi 10.1038/Nature06159.

Conaway, R. C., and J. W. Conaway. 2011. "Origins and activity of the Mediator complex." *Seminars in Cell & Developmental Biology* 22 (7):729-734. doi: Doi 10.1016/J.Semcdb.2011.07.021.

Core, L. J., J. J. Waterfall, and J. T. Lis. 2008. "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." *Science* 322 (5909):1845-8. doi: 1162228 [pii] 10.1126/science.1162228.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 2010. "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50):21931-21936. doi: Doi 10.1073/Pnas.1016071107.

Davis, R. L., H. Weintraub, and A. B. Lassar. 1987. "Expression of a Single Transfected Cdna Converts Fibroblasts to Myoblasts." *Cell* 51 (6):987-1000. doi: Doi 10.1016/0092-8674(87)90585-X.

Decker, K. B., and D. M. Hinton. 2013. "Transcription Regulation at the Core: Similarities Among Bacterial, Archaeal, and Eukaryotic RNA Polymerases." *Annual Review of Microbiology, Vol 67* 67:113-139. doi: Doi 10.1146/Annurev-Micro-092412-155756.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485 (7398):376-80. doi: nature11082 [pii] 10.1038/nature11082.

Dowen, J. M., Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schuijers, T. I. Lee, K. Zhao, and R. A. Young. 2014. "Control of cell identity genes occurs in insulated neighborhoods in Mammalian chromosomes." *Cell* 159 (2):374-87. doi: 10.1016/j.cell.2014.09.030.

Dynan, W. S., and R. Tjian. 1983. "The Promoter-Specific Transcription Factor-Sp1 Binds to Upstream Sequences in the Sv40 Early Promoter." *Cell* 35 (1):79-87. doi: Doi 10.1016/0092-8674(83)90210-6.

Eberhardy, S. R., and P. J. Farnham. 2002. "Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter." *J Biol Chem* 277 (42):40156-62. doi: 10.1074/jbc.M207441200.

Ebmeier, C. C., and D. J. Taatjes. 2010. "Activator-Mediator binding regulates Mediator-cofactor interactions." *Proc Natl Acad Sci U S A* 107 (25):11283-8. doi: 10.1073/pnas.0914215107.

Euskirchen, G. M., R. K. Auerbach, E. Davidov, T. A. Gianoulis, G. N. Zhong, J. Rozowsky, N. Bhardwaj, M. B. Gerstein, and M. Snyder. 2011. "Diverse Roles and Interactions of the SWI/SNF Chromatin Remodeling Complex Revealed Using Global Approaches." *Plos Genetics* 7 (3). doi: Artn E1002008 Doi 10.1371/Journal.Pgen.1002008.

Factor, D. C., O. Corradin, G. E. Zentner, A. Saiakhova, L. Y. Song, J. G. Chenoweth, R. D. McKay, G. E. Crawford, P. C. Scacheri, and P. J. Tesar. 2014. "Epigenomic Comparison Reveals Activation of "Seed" Enhancers during Transition from Naive to Primed Pluripotency." *Cell Stem Cell* 14 (6):854-863. doi: Doi 10.1016/J.Stem.2014.05.005.

Feng, Q., H. B. Wang, H. H. Ng, H. Erdjument-Bromage, P. Tempst, K. Struhl, and Y. Zhang. 2002. "Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain." *Current Biology* 12 (12):1052-1058. doi: Pii S0960-9822(02)00901-6 Doi 10.1016/S0960-9822(02)00901-6.

Filippova, G. N., S. Fagerlie, E. M. Klenova, C. Myers, Y. Dehner, G. Goodwin, P. E. Neiman, S. J. Collins, and V. V. Lobanenkov. 1996. "An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes." *Molecular and Cellular Biology* 16 (6):2802-2813.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. 2009. "An oestrogen-receptor-alpha-bound human chromatin interactome." *Nature* 462 (7269):58-64. doi: nature08497 [pii] 10.1038/nature08497.

Gaszner, M., J. Vazquez, and P. Schedl. 1999. "The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction." *Genes & Development* 13 (16):2098-2107. doi: DOI 10.1101/gad.13.16.2098.

Geyer, P. K., and V. G. Corces. 1992. "DNA Position-Specific Repression of Transcription by a Drosophila Zinc Finger Protein." *Genes & Development* 6 (10):1865-1873. doi: DOI 10.1101/gad.6.10.1865.

Gibcus, J. H., and J. Dekker. 2013. "The hierarchy of the 3D genome." *Mol Cell* 49 (5):773-82. doi: S1097-2765(13)00139-1 [pii] 10.1016/j.molcel.2013.02.011.

Gilchrist, D. A. 2010. "Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation." *Cell* 143:540-551.

Gordan, R., N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk. 2013. "Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape." *Cell Reports* 3 (4):1093-1104. doi: Doi 10.1016/J.Celrep.2013.03.014.

Gorkin, D. U., D. Leung, and B. Ren. 2014. "The 3D Genome in Transcriptional Regulation and Pluripotency." *Cell Stem Cell* 14 (6):762-775. doi: DOI 10.1016/j.stem.2014.05.017.

Graf, T., and T. Enver. 2009. "Forcing cells to change lineages." *Nature* 462 (7273):587-94. doi: nature08533 [pii] 10.1038/nature08533.

Grunberg, S., and S. Hahn. 2013. "Structural insights into transcription initiation by RNA polymerase II." *Trends Biochem. Sci.* 38:603-611.

Guenther, M. G., L. N. Lawton, T. Rozovskaia, G. M. Frampton, S. S. Levine, T. L. Volkert, C. M. Croce, T. Nakamura, E. Canaani, and R. A. Young. 2008. "Aberrant chromatin at genes encoding stem cell regulators in human mixed-

lineage leukemia." *Genes Dev* 22 (24):3403-8. doi: 22/24/3403 [pii] 10.1101/gad.1741408.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. 2007. "A chromatin landmark and transcription initiation at most promoters in human cells." *Cell* 130 (1):77-88. doi: S0092-8674(07)00681-2 [pii] 10.1016/j.cell.2007.05.042.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. 2011. "CTCF-mediated functional chromatin interactome in pluripotent cells." *Nat Genet* 43 (7):630-8. doi: ng.857 [pii] 10.1038/ng.857.

Hargreaves, D. C., and G. R. Crabtree. 2011. "ATP-dependent chromatin remodeling: genetics, genomics and mechanisms." *Cell Research* 21 (3):396-420. doi: DOI 10.1038/cr.2011.32.

Heintzman, N. D., G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren. 2009. "Histone modifications at human enhancers reflect global cell-type-specific gene expression." *Nature* 459 (7243):108-12. doi: nature07829 [pii] 10.1038/nature07829.

Heinz, S., C. E. Romanoski, C. Benner, and C. K. Glass. 2015. "The selection and function of cell type-specific enhancers." *Nat Rev Mol Cell Biol.* doi: 10.1038/nrm3949.

Hendrix, D. A., J. W. Hong, J. Zeitlinger, D. S. Rokhsar, and M. S. Levine. 2008. "Promoter elements associated with RNA Pol II stalling in the Drosophila embryo." *Proc Natl Acad Sci U S A* 105 (22):7762-7. doi: 10.1073/pnas.0802406105.

Hsin, J. P., and J. L. Manley. 2012. "The RNA polymerase II CTD coordinates transcription and RNA processing." *Genes & Development* 26 (19):2119-2137. doi: Doi 10.1101/Gad.200303.112.

Jacob, F., and J. Monod. 1961. "Genetic Regulatory Mechanisms in Synthesis of Proteins." *Journal of Molecular Biology* 3 (3):318-&. doi: Doi 10.1016/S0022-2836(61)80072-7.

Jolma, A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. H. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. 2013. "DNA-Binding Specificities of Human Transcription Factors." *Cell* 152 (1-2):327-339. doi: DOI 10.1016/j.cell.2012.12.009.

Joshi, R., J. M. Passner, R. Rohs, R. Jain, A. Sosinsky, M. A. Crickmore, V. Jacob, A. K. Aggarwal, B. Honig, and R. S. Mann. 2007. "Functional specificity of a Hox protein mediated by the recognition of minor groove structure." *Cell* 131 (3):530-543. doi: Doi 10.1016/J.Cell.2007.09.024.

Juven-Gershon, T., J. Y. Hsu, J. W. M. Theisen, and J. T. Kadonaga. 2008. "The RNA polymerase II core promoter - the gateway to transcription." *Current Opinion in Cell Biology* 20 (3):253-259. doi: DOI 10.1016/j.ceb.2008.03.003.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. 2010. "Mediator and cohesin connect gene expression and chromatin architecture." *Nature* 467 (7314):430-5. doi: 10.1038/nature09380.

Kellum, R., and P. Schedl. 1991. "A Position-Effect Assay for Boundaries of Higher-Order Chromosomal Domains." *Cell* 64 (5):941-950. doi: Doi 10.1016/0092-8674(91)90318-S.

Kim, J., J. Chu, X. Shen, J. Wang, and S. H. Orkin. 2008. "An extended transcriptional network for pluripotency of embryonic stem cells." *Cell* 132 (6):1049-61. doi: S0092-8674(08)00328-0 [pii] 10.1016/j.cell.2008.02.039.

Kim, T. K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. 2010. "Widespread transcription at neuronal activity-regulated enhancers." *Nature* 465 (7295):182-U65. doi: DOI 10.1038/nature09033.

Klenova, E. M., R. H. Nicolas, H. F. Paterson, A. F. Carne, C. M. Heath, G. H. Goodwin, P. E. Neiman, and V. V. Lobanenkov. 1993. "Ctcf, a Conserved Nuclear Factor Required for Optimal Transcriptional Activity of the Chicken C-Myc Gene, Is an 11-Zn-Finger Protein Differentially Expressed in Multiple Forms." *Molecular and Cellular Biology* 13 (12):7612-7624.

Kornberg, R. D. 1974. "Chromatin structure: a repeating unit of histones and DNA." *Science* 184 (4139):868-71.

Kornberg, R. D. 2005. "Mediator and the mechanism of transcriptional activation." *Trends in Biochemical Sciences* 30 (5):235-239. doi: Doi 10.1016/J.Tibs.2005.03.011.

Kornberg, R. D. 2007. "The molecular basis of eukaryotic transcription." *Proc Natl Acad Sci U S A* 104 (32):12955-61. doi: 10.1073/pnas.0704138104.

Kornberg, R. D., and Y. Lorch. 1999. "Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome." *Cell* 98 (3):285-94.

Kouzarides, T. 2007. "Chromatin modifications and their function." *Cell* 128 (4):693-705. doi: DOI 10.1016/j.cell.2007.02.005.

Kurukuti, S., V. K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. H. Zhao, V. Lobanenkov, W. Reik, and R. Ohlsson. 2006. "CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2." *Proceedings of the National Academy of Sciences of the United States of America* 103 (28):10684-10689. doi: DOI 10.1073/pnas.0600326103.

Lachner, M., N. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. 2001. "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins." *Nature* 410 (6824):116-120. doi: Doi 10.1038/35065132.

Lassar, A. B., B. M. Paterson, and H. Weintraub. 1986. "Transfection of a DNA Locus That Mediates the Conversion of 10t1/2 Fibroblasts to Myoblasts." *Cell* 47 (5):649-656. doi: Doi 10.1016/0092-8674(86)90507-6.

Lee, T. I., R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. 2006. "Control of developmental regulators by Polycomb in human embryonic stem cells." *Cell* 125 (2):301-13. doi: S0092-8674(06)00384-9 [pii] 10.1016/j.cell.2006.02.043.

Lee, T. I., and R. A. Young. 2000. "Transcription of eukaryotic protein-coding genes." *Annual Review of Genetics* 34:77-137. doi: Doi 10.1146/Annurev.Genet.34.1.77.

Lee, T. I., and R. A. Young. 2013. "Transcriptional regulation and its misregulation in disease." *Cell* 152 (6):1237-51. doi: 10.1016/j.cell.2013.02.014.

Lenhard, B., A. Sandelin, and P. Carninci. 2012. "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." *Nature Rev. Genet.* 13:233-245.

Lettice, L. A., S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. 2003. "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly." *Human Molecular Genetics* 12 (14):1725-1735. doi: Doi 10.1093/Hmg/Ddg180.

Levine, M. 2010. "Transcriptional enhancers in animal development and evolution." *Curr. Biol.* 20:R754-R763.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. 2009. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326 (5950):289-93. doi: 326/5950/289 [pii] 10.1126/science.1181369.

Liu, Z., D. R. Scannell, M. B. Eisen, and R. Tjian. 2011. "Control of Embryonic Stem Cell Lineage Commitment by Core Promoter Factor, TAF3." *Cell* 146 (5):720-731. doi: Doi 10.1016/J.Cell.2011.08.005.

Lobanenkov, V. V., V. V. Adler, E. M. Klenova, R. H. Nicolas, and G. H. Goodwin. 1990. "Ccctc-Binding Factor (Ctcf) - a Novel Sequence-Specific DNA-Binding Protein Which Interacts with the 5'-Flanking Sequence of the Chicken C-Myc Gene." *Gene Regulation and Aids : Transcriptional Activation, Retroviruses, and Pathogenesis* 7:45-68.

Lu, H., O. Flores, R. Weinmann, and D. Reinberg. 1991. "The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex." *Proc Natl Acad Sci U S A* 88 (22):10004-8.

Lu, H., L. Zawel, L. Fisher, J. M. Egly, and D. Reinberg. 1992. "Human General Transcription Factor-Iih Phosphorylates the C-Terminal Domain of Rna Polymerase-Ii." *Nature* 358 (6388):641-645. doi: Doi 10.1038/358641a0.

Majumder, P., J. A. Gomez, B. P. Chadwick, and J. M. Boss. 2008. "The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions." *Journal of Experimental Medicine* 205 (4):785-798. doi: Doi 10.1084/Jem.20071843.

Malik, S., and R. G. Roeder. 2005. "Dynamic regulation of pol II transcription by the mammalian Mediator complex." *Trends Biochem Sci* 30 (5):256-63.

Maniatis, T., J. V. Falvo, T. H. Kim, T. K. Kim, C. H. Lin, B. S. Parekh, and M. G. Wathelet. 1998. "Structure and function of the interferon-beta enhanceosome." *Cold Spring Harbor Symposia on Quantitative Biology* 63:609-620. doi: Doi 10.1101/Sqb.1998.63.609.

Marshall, N. F., J. Peng, Z. Xie, and D. H. Price. 1996. "Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase." *J Biol Chem* 271 (43):27176-83.

Marshall, N. F., and D. H. Price. 1992. "Control of formation of two distinct classes of RNA polymerase II elongation complexes." *Mol Cell Biol* 12 (5):2078-90.

Marshall, N. F., and D. H. Price. 1995. "Purification of P-TEFb, a transcription factor required for the transition into productive elongation." *J Biol Chem* 270 (21):12335-8.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young. 2008. "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." *Cell* 134 (3):521-33. doi: S0092-8674(08)00938-0 [pii] 10.1016/j.cell.2008.07.020.

Masui, S., Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. H. Ko, and H. Niwa. 2007. "Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells." *Nature Cell Biology* 9 (6):625-U26. doi: Doi 10.1038/Ncb1589.

McCracken, S., N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Poster, S. Shuman, D. L. Bentley, and Amgen EST Program. 1997. "5 '-capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II." *Genes & Development* 11 (24):3306-3318. doi: Doi 10.1101/Gad.11.24.3306.

Merkenschlager, M., and D. T. Odom. 2013. "CTCF and cohesin: linking gene regulatory elements with their targets." *Cell* 152 (6):1285-97. doi: S0092-8674(13)00218-3 [pii] 10.1016/j.cell.2013.02.029.

Misteli, T. 2007. "Beyond the sequence: Cellular organization of genome function." *Cell* 128 (4):787-800. doi: DOI 10.1016/j.cell.2007.01.028.

Moon, H., G. Filippova, D. Loukinov, E. Pugacheva, Q. Chen, S. T. Smith, A. Munhall, B. Grewe, M. Bartkuhn, R. Arnold, L. J. Burke, R. Renkawitz-Pohl, R. Ohlsson, J. M. Zhou, R. Renkawitz, and V. Lobanenkov. 2005. "CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator." *Embo Reports* 6 (2):165-170. doi: Doi 10.1038/Sj.Embor.7400334.

Moore, M. J., and N. J. Proudfoot. 2009. "Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation." *Cell* 136 (4):688-700. doi: DOI 10.1016/j.cell.2009.02.001.

Morris, S. A., S. Baek, M. H. Sung, S. John, M. Wiench, T. A. Johnson, R. L. Schiltz, and G. L. Hager. 2014. "Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions." *Nature Structural & Molecular Biology* 21 (1):73-+. doi: Doi 10.1038/Nsmb.2718.

Mullen, A. C., D. A. Orlando, J. J. Newman, J. Loven, R. M. Kumar, S. Bilodeau, J. Reddy, M. G. Guenther, R. P. DeKoter, and R. A. Young. 2011. "Master transcription factors determine cell-type-specific responses to TGF-beta signaling." *Cell* 147 (3):565-76. doi: S0092-8674(11)01134-2 [pii] 10.1016/j.cell.2011.08.050.

Muse, G. W. 2007. "RNA polymerase is poised for activation across the genome." *Nature Genet.* 39:1507-1511.

Myer, V. E., and R. A. Young. 1998. "RNA polymerase II holoenzymes and subcomplexes." *Journal of Biological Chemistry* 273 (43):27757-27760. doi: DOI 10.1074/jbc.273.43.27757.

Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen,

J. Dekker, and E. Heard. 2012. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." *Nature* 485 (7398):381-5. doi: nature11049 [pii]

10.1038/nature11049.

Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, G. M. Frampton, A. C. Drake, I. Leskov, B. Nilsson, F. Preffer, D. Dombkowski, J. W. Evans, T. Liefeld, J. S. Smutko, J. Chen, N. Friedman, R. A. Young, T. R. Golub, A. Regev, and B. L. Ebert. 2011. "Densely interconnected transcriptional circuits control cell states in human hematopoiesis." *Cell* 144 (2):296-309. doi: S0092-8674(11)00005-5 [pii] 10.1016/j.cell.2011.01.004.

Odom, D. T., R. D. Dowell, E. S. Jacobsen, L. Nekludova, P. A. Rolfe, T. W. Danford, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. 2006. "Core transcriptional regulatory circuitry in human hepatocytes." *Mol Syst Biol* 2:2006 0017. doi: msb4100059 [pii] 10.1038/msb4100059.

Ohlsson, R., R. Renkawitz, and V. Lobanenkov. 2001. "CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease." *Trends in Genetics* 17 (9):520-527. doi: Doi 10.1016/S0168-9525(01)02366-6.

Ong, C. T., and V. G. Corces. 2011. "Enhancer function: new insights into the regulation of tissue-specific gene expression." *Nature Rev. Genet.* 12:283-293.

Orkin, S. H., and K. Hochedlinger. 2011. "Chromatin connections to pluripotency and cellular reprogramming." *Cell* 145 (6):835-50. doi: 10.1016/j.cell.2011.05.019.

Oudet, P., M. Grossbellard, and P. Chambon. 1975. "Electron-Microscopic and Biochemical Evidence That Chromatin Structure Is a Repeating Unit." *Cell* 4 (4):281-300. doi: Doi 10.1016/0092-8674(75)90149-X.

Peterlin, B. M., and D. H. Price. 2006. "Controlling the elongation phase of transcription with P-TEFb." *Molecular Cell* 23 (3):297-305. doi: DOI 10.1016/j.molcel.2006.06.014.

Phillips, J. E., and V. G. Corces. 2009. "CTCF: master weaver of the genome." *Cell* 137 (7):1194-211. doi: S0092-8674(09)00699-0 [pii] 10.1016/j.cell.2009.06.001.

Phillips-Cremins, J. E., M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces. 2013. "Architectural protein subclasses shape 3D organization of genomes during lineage commitment." *Cell* 153 (6):1281-95. doi: S0092-8674(13)00529-1 [pii] 10.1016/j.cell.2013.04.053.

Pokholok, D. K., C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young. 2005. "Genome-wide map of nucleosome acetylation and methylation in yeast." *Cell* 122 (4):517-27. doi: 10.1016/j.cell.2005.06.026.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. "A unique chromatin signature uncovers early developmental enhancers in humans." *Nature* 470 (7333):279-+. doi: Doi 10.1038/Nature09692.

Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. 2010. "c-Myc regulates transcriptional pause release." *Cell* 141 (3):432-45. doi: S0092-8674(10)00318-1 [pii] 10.1016/j.cell.2010.03.030.

Ramskold, D., E. T. Wang, C. B. Burge, and R. Sandberg. 2009. "An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data." *Plos Computational Biology* 5 (12). doi: ARTN e1000598 DOI 10.1371/journal.pcbi.1000598.

Rivera, C. M., and B. Ren. 2013. "Mapping Human Epigenomes." *Cell* 155 (1):39-55. doi: DOI 10.1016/j.cell.2013.09.011.

Roeder, R. G. 1996. "The role of general initiation factors in transcription by RNA polymerase II." *Trends Biochem Sci* 21 (9):327-35.

Roeder, R. G. 2005. "Transcriptional regulation and the role of diverse coactivators in animal cells." *FEBS Lett* 579 (4):909-15. doi: S0014-5793(04)01531-5 [pii] 10.1016/j.febslet.2004.12.007.

Rohs, R., X. S. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. 2010. "Origins of Specificity in Protein-DNA Recognition." *Annual Review of Biochemistry, Vol 79* 79:233-269. doi: DOI 10.1146/annurev-biochem-060408-091030.

Rohs, R., S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. 2009. "The role of DNA shape in protein-DNA recognition." *Nature* 461 (7268):1248-U81. doi: Doi 10.1038/Nature08473.

Roy, A. L., and D. S. Singer. 2015. "Core promoters in transcription: old problem, new insights." *Trends in Biochemical Sciences* 40 (3):165-171. doi: Doi 10.1016/J.Tibs.2015.01.007.

Sainsbury, S., C. Bernecky, and P. Cramer. 2015. "Structural basis of transcription initiation by RNA polymerase II." *Nature Reviews Molecular Cell Biology* 16 (3):129-143. doi: DOI 10.1038/nrm3952.

Sanda, T., L. N. Lawton, M. I. Barrasa, Z. P. Fan, H. Kohlhammer, A. Gutierrez, W. Ma, J. Tatarek, Y. Ahn, M. A. Kelliher, C. H. Jamieson, L. M. Staudt, R. A. Young, and A. T. Look. 2012. "Core transcriptional regulatory circuit controlled by

the TAL1 complex in human T cell acute lymphoblastic leukemia." *Cancer Cell* 22 (2):209-21. doi: S1535-6108(12)00256-5 [pii] 10.1016/j.ccr.2012.06.007.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker. 2012. "The long-range interaction landscape of gene promoters." *Nature* 489 (7414):109-13. doi: nature11279 [pii] 10.1038/nature11279.

Segal, E., T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. 2008. "Predicting expression patterns from regulatory sequence in Drosophila segmentation." *Nature* 451 (7178):535-U1. doi: Doi 10.1038/Nature06496.

Seitan, V. C., M. S. Krangel, and M. Merkenschlager. 2012. "Cohesin, CTCF and lymphocyte antigen receptor locus rearrangement." *Trends in Immunology* 33 (4):153-159. doi: Doi 10.1016/J.It.2012.02.004.

Shikama, N., J. Lyon, and N. B. LaThangue. 1997. "The p300/CBP family: Integrating signals with transcription factors and chromatin." *Trends in Cell Biology* 7 (6):230-236.

Slattery, M., T. Y. Zhou, L. Yang, A. C. D. Machado, R. Gordan, and R. Rohs. 2014. "Absence of a simple code: how transcription factors read the genome." *Trends in Biochemical Sciences* 39 (9):381-399. doi: DOI 10.1016/j.tibs.2014.07.002.

Smale, S. T., and J. T. Kadonaga. 2003. "The RNA polymerase II core promoter." *Annual Review of Biochemistry* 72:449-479. doi: DOI 10.1146/annurev.biochem.72.121801.161520.

Sofueva, S., E. Yaffe, W. C. Chan, D. Georgopoulou, M. Vietri Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur. 2013. "Cohesin-mediated interactions organize chromosomal domain architecture." *EMBO J.* doi: 10.1038/emboj.2013.237.

Spitz, F., and E. E. Furlong. 2012. "Transcription factors: from enhancer binding to developmental control." *Nat Rev Genet* 13 (9):613-26. doi: nrg3207 [pii] 10.1038/nrg3207.

Stella, S., D. Cascio, and R. C. Johnson. 2010. "The shape of the DNA minor groove directs binding by the DNA-bending protein Fis." *Genes Dev* 24 (8):814-26. doi: 10.1101/gad.1900610.

Stormo, G. D., and Y. Zhao. 2010. "Determining the specificity of protein-DNA interactions." *Nature Reviews Genetics* 11 (11):751-760. doi: DOI 10.1038/nrg2845.

Sun, F. L., and S. C. R. Elgin. 1999. "Putting boundaries on silence." *Cell* 99 (5):459-462. doi: Doi 10.1016/S0092-8674(00)81534-2.

Sun, X. J., J. Wei, X. Y. Wu, M. Hu, L. Wang, H. H. Wang, Q. H. Zhang, S. J. Chen, Q. H. Huang, and Z. Chen. 2005. "Identification and characterization of a

novel human histone H3 lysine 36-specific methyltransferase." *Journal of Biological Chemistry* 280 (42):35261-35271. doi: DOI 10.1074/jbc.M504012200.

Taatjes, D. J. 2010. "The human Mediator complex: a versatile, genome-wide regulator of transcription." *Trends in Biochemical Sciences* 35 (6):315-322. doi: Doi 10.1016/J.Tibs.2010.02.004.

Takahashi, H., T. J. Parmely, S. Sato, C. Tomomori-Sato, C. A. Banks, S. E. Kong, H. Szutorisz, S. K. Swanson, S. Martin-Brown, M. P. Washburn, L. Florens, C. W. Seidel, C. Lin, E. R. Smith, A. Shilatifard, R. C. Conaway, and J. W. Conaway. 2011. "Human mediator subunit MED26 functions as a docking site for transcription elongation factors." *Cell* 146 (1):92-104. doi: 10.1016/j.cell.2011.06.005.

Takahashi, K., and S. Yamanaka. 2006. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." *Cell* 126 (4):663-76. doi: S0092-8674(06)00976-7 [pii] 10.1016/j.cell.2006.07.024.

Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. 2002. "Looping and interaction between hypersensitive sites in the active beta-globin locus." *Mol Cell* 10 (6):1453-65. doi: S1097276502007815 [pii].

Trompouki, E., T. V. Bowman, L. N. Lawton, Z. P. Fan, D. C. Wu, A. DiBiase, C. S. Martin, J. N. Cech, A. K. Sessa, J. L. Leblanc, P. L. Li, E. M. Durand, C. Mosimann, G. C. Heffner, G. Q. Daley, R. F. Paulson, R. A. Young, and L. I. Zon. 2011. "Lineage Regulators Direct BMP and Wnt Pathways to Cell-Specific Programs during Differentiation and Regeneration." *Cell* 147 (3):577-589. doi: Doi 10.1016/J.Cell.2011.09.044.

Vakoc, C. R., D. L. Letting, N. Gheldof, T. Sawado, M. A. Bender, M. Groudine, M. J. Weiss, J. Dekker, and G. A. Blobel. 2005. "Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1." *Mol Cell* 17 (3):453-62. doi: S1097276505010154 [pii] 10.1016/j.molcel.2004.12.028.

Vannini, A., and P. Cramer. 2012. "Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries." *Molecular Cell* 45 (4):439-446. doi: DOI 10.1016/j.molcel.2012.01.023.

Wada, T., T. Takagi, Y. Yamaguchi, D. Watanabe, and H. Handa. 1998. "Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro." *Embo Journal* 17 (24):7395-7403. doi: DOI 10.1093/emboj/17.24.7395.

Wang, Z. B., C. Z. Zang, K. R. Cui, D. E. Schones, A. Barski, W. Q. Peng, and K. J. Zhao. 2009. "Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes." *Cell* 138 (5):1019-1031. doi: Doi 10.1016/J.Cell.2009.06.049.

Wang, Z., E. Oron, B. Nelson, S. Razis, and N. Ivanova. 2012. "Distinct Lineage Specification Roles for NANOG, OCT4, and SOX2 in Human Embryonic Stem Cells." *Cell Stem Cell* 10 (4):440-454. doi: Doi 10.1016/J.Stem.2012.02.016.

Weiss, M. J., and S. H. Orkin. 1995. "Transcription Factor Gata-1 Permits Survival and Maturation of Erythroid Precursors by Preventing Apoptosis." *Proceedings of the National Academy of Sciences of the United States of America* 92 (21):9623-9627. doi: Doi 10.1073/Pnas.92.21.9623.

West, A. G., M. Gaszner, and G. Felsenfeld. 2002. "Insulators: many functions, many mechanisms." *Genes & Development* 16 (3):271-288. doi: DOI 10.1101/gad.954702.

Wu, C. H., Y. Yamaguchi, L. R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. 2003. "NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila." *Genes Dev* 17 (11):1402-14. doi: 10.1101/gad.1091403.

Xiao, T. J., J. Wallace, and G. Felsenfeld. 2011. "Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity." *Molecular and Cellular Biology* 31 (11):2174-2183. doi: Doi 10.1128/Mcb.05093-11.

Yamaguchi, Y., T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. 1999. "NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation." *Cell* 97 (1):41-51. doi: S0092-8674(00)80713-8 [pii].

Yan, J., M. Enge, T. Whitington, K. Dave, J. Liu, I. Sur, B. Schmierer, A. Jolma, T. Kivioja, M. Taipale, and J. Taipale. 2013. "Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites." *Cell* 154 (4):801-13. doi: 10.1016/j.cell.2013.07.034.

Yin, J. W., and G. Wang. 2014. "The Mediator complex: a master coordinator of transcription and cell lineage development." *Development* 141 (5):977-987. doi: DOI 10.1242/dev.098392.

Young, R. A. 2011. "Control of the embryonic stem cell state." *Cell* 144 (6):940-54. doi: 10.1016/j.cell.2011.01.032.

Zaret, K. S., and J. S. Carroll. 2011. "Pioneer transcription factors: establishing competence for gene expression." *Genes Dev.* 25:2227-2241.

Zeitlinger, J. 2007. "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." *Nature Genet.* 39:1512-1516.

Zhao, K., C. M. Hart, and U. K. Laemmli. 1995. "Visualization of Chromosomal Domains with Boundary Element-Associated Factor Beaf-32." *Cell* 81 (6):879-889. doi: Doi 10.1016/0092-8674(95)90008-X.

Zuin, J., J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, W. F. van Ijcken, F. G. Grosveld, B. Ren, and K. S. Wendt. 2014. "Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells." *Proc Natl Acad Sci U S A* 111 (3):996-1001. doi: 10.1073/pnas.1317788111.

Chapter 2

Functional retinal pigment epithelium-like cells from human fibroblasts

Ana C. D'Alessio[1,6], Zi Peng Fan[1,2,6], Katherine J. Wert[1], Malkiel A. Cohen[1], Janmeet S. Saini[4,5], Evan Cohick[1], Carol Charniga[4], Daniel Dadon[1,3], Nancy M. Hannett[1], Sally Temple[4], Rudolf Jaenisch[1,3], Tong Ihn Lee[1], Richard A. Young[1,3]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142

[2]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

[4]Neural Stem Cell Institute, Rensselaer, NY 12144

[5]Department of Biomedical Sciences, University at Albany, Albany, NY 12201

[6]These authors contributed equally

Personal Contribution to the Project

This work was a close collaboration between Ana C. D'Alessio, Tong Ihn Lee and myself. I performed all the computational analyses. Ana C. D'Alessio, Katherine J. Wert, Malkiel A. Cohen, Evan Cohick, and Nancy M. Hannett performed the experiments. The manuscript was written by Ana C. D'Alessio, Tong Ihn Lee, Richard A. Young, and myself.

SUMMARY

The retinal pigment epithelium (RPE) provides vital support to photoreceptor cells and its dysfunction is associated with the onset and progression of age-related macular degeneration (AMD). Surgical provision of RPE cells may ameliorate AMD and thus it may be valuable to develop sources of patient-matched RPE cells via reprogramming. We used a computational approach to generate an atlas of candidate master transcriptional regulators for a broad spectrum of human cells and then used candidate RPE regulators to guide investigation of the transcriptional regulatory circuitry of RPE cells and to reprogram human fibroblasts into RPE-like cells. The RPE-like cells share key features with RPE cells derived from healthy individuals, including morphology, gene expression and function. The approach described here should be useful for systematically discovering regulatory circuitries and reprogramming cells for additional clinically important cell types.

INTRODUCTION

The retinal pigment epithelium (RPE) provides vital support to photoreceptor cells in the vertebrate eye (Strauss 2005, Sparrow, Hicks, and Hamel 2010). Progressive degeneration of the retinal pigment epithelium is a major cause of age-related macular degeneration (AMD), which affects nearly 20% of individuals in aging populations (Lim et al. 2012). Surgical provision of healthy RPE cells has been used with some success in individuals with AMD (Binder et al. 2007, da Cruz et al. 2007) and there is considerable interest in generating patient-matched RPE cells for regenerative therapy. Human embryonic stem cell (ESC)-derived RPE cells have been transplanted into patients with AMD and initial results suggest visual improvement with no rejection or adverse outcomes (Schwartz et al. 2012, Schwartz et al. 2014). Several clinical trials are currently assessing the use of RPE cells in the treatment of ocular disorders (Cyranoski 2013, 2014)(Clinical trials.gov NCT01674829, NCT01345006, NCT01344993, NCT01625559, NCT01469832). The RPE cells being used for these clinical trials are differentiated from human ESC or induced pluripotent stem cell (iPSC) lines (Kamao et al. 2014).

The potential of RPE cells for regenerative medicine has led to interest in the possibility that RPE cells might be obtained by direct reprogramming from fibroblasts, which is an alternative to the use of stem-cell-differentiated cells for cell-based replacement therapies. For some cell types, direct reprogramming can be achieved by ectopic expression of key transcription factors of the target cell type in cells of a different type (Vierbuchen and Wernig 2012, Buganim, Faddah,

and Jaenisch 2013, Morris and Daley 2013, Sancho-Martinez, Baek, and Izpisua Belmonte 2012, Yamanaka 2012, Graf and Enver 2009). Due to limited knowledge of the key factors for each cell type, which we will henceforth call master transcription factors, it is not currently possible to obtain various clinically relevant cell types by this approach. It would be valuable to identify candidate master transcription factors for all cell types: an atlas of such regulators would complement ENCODEs encyclopedia of DNA elements (Stergachis et al. 2013, Rivera and Ren 2013), could guide exploration of the core transcriptional regulatory circuitry of cells (Young 2011), and will enable more systematic research into the mechanistic and global functions of these key regulators of cell identity (Soufi, Donahue, and Zaret 2012, Xie and Ren 2013, Iwafuchi-Doi and Zaret 2014, Henriques et al. 2013). The identification of master transcription factors in all cell types should also facilitate advances in direct reprogramming for clinically relevant cell types, including RPE cells.

We describe here the identification of candidate master transcriptional factors for a broad spectrum of human cells and the use of predicted RPE factors to investigate the transcriptional regulatory circuitry of RPE cells and to reprogram human fibroblasts into RPE-like cells. A novel computational approach was used to systematically identify candidate master transcription factors for most known human cell types. Genetic perturbation and genome-wide binding profiles of the predicted RPE master transcription factors confirmed the importance of these factors for RPE cell identity and produced a model of RPE core regulatory circuitry. Ectopic expression of predicted RPE master

transcription factors in human fibroblasts produced cells that share key features with RPE cells derived from healthy individuals, including morphology, gene expression and function. These results suggest that the atlas of candidate master transcription factors should be useful for systematically discovering regulatory circuitries for many cell types and for reprogramming additional clinically important cell types.

RESULTS

Candidate master transcription factors for human cells

The master transcription factors (TFs) that are known to be important for establishment or maintenance of cell state, and that are components of most successful reprogramming factor cocktails, are expressed at high levels in specific cell types (Lee and Young 2013). A computational approach was developed that exploits this feature to identify candidate master TFs in all cell types for which gene expression data is available (Figure 1A). The algorithm quantifies both the relative level and the cell-type-specificity of gene expression by using an entropy-based measure of Jensen-Shannon divergence (Cabili et al. 2011) to compare the expression of a transcription factor in a cell type of interest to the expression of that factor across a range of cell types (Extended Experimental Procedures). The algorithm assumes an idealized case where a transcription factor is expressed to a high level in a single cell type and not expressed in any other cell type, then generates a specificity score based on how well the actual data matches with this idealized case, and ranks each

transcription factor accordingly. This approach has additional features that make it flexible yet robust. It is modular and expandable to the expression profiles of disparate cell types from different laboratories. Multiple expression profiles of a query cell type can be used to increase the robustness of the predictions. The algorithm also takes advantage of the multiplicity of expression profiles to favor those gene probes that are ranked highly and consistently across multiple profiles.

This approach was used to predict master TFs for 106 cell types/tissues represented in the Human Body Index collection of expression data together with some additional well-studied cell types (Figure 1B, Table S1, Table S2, Extended Experimental Procedures). Because embryonic stem cells (ESCs) are among the best-characterized cells, ESCs represented a useful first test case for the approach. The top-ranked factors for embryonic stem cells included the reprogramming factors OCT4/POU5F1, SOX2, NANOG, SALL4 and MYCN and additional factors known to be important for ESCs (ZIC2, ZIC3, OTX2, ZSCAN10)(Figure 1C) (Avilion et al. 2003, Boyer et al. 2005, Chambers et al. 2003, Ivanova et al. 2006, Kim et al. 2008, Wang, Kueh, et al. 2007, Wang, Teh, et al. 2007). The top ranked factors for other well-studied cell types included the transcription factors that have been shown to be capable of trans-differentiating fibroblasts into various other cell types (Table S2). Thus, this compendium of candidate master TFs should prove to be a useful resource for future studies of transcriptional regulatory networks and perhaps for reprogramming cell state.

RPE master transcription factors, super-enhancers and core circuitry

To improve our understanding of the transcriptional control of RPE cells, we carried out a study of the candidate master TFs identified for these cells (Figure 1D). We selected nine top scoring transcription factors - PAX6, LHX2, OTX2, SOX9, MITF, SIX3, ZNF92, GLIS3, and FOXD1 - for further study. Among these, PAX6, OTX2 and MITF have previously been implicated in retinal pigment cell development (Bharti et al. 2012, Martinez-Morales et al. 2003, Matsuo et al. 1995), and SOX9 has been shown to interact with OTX2 and MITF (Martinez-Morales et al. 2003, Masuda and Esumi 2010). Furthermore, PAX6, OTX2, MITF and five other TFs (MYC, KLF4, NRL, CRX and RAX) have been shown to induce an RPE-like progenitor state in fibroblasts (Zhang et al. 2014).

Well-studied master TFs are essential for maintenance of the gene expression program that controls cell identity, so we determined whether the RPE master TF candidates are essential for maintenance of the RPE gene expression program. We successfully knocked-down expression of eight (PAX6, OTX2, SOX9, MITF, SIX3, ZNF92, GLIS3 and FOXD1) of the nine candidate factors in human RPE cells (Figure 2A, Table S3). Efficient knockdown of LHX2 was not successful, despite multiple attempts with several shRNA constructs. For each the eight TFs where efficient knockdown was achieved, reduced levels of the TF mRNA led to reduced expression of three well-studied genes known to be key to RPE function: RPE65, CRALBP and TYP (Figure 2B). RPE65 and CRALBP encode two proteins that function in the visual cycle and TYR encodes an enzyme responsible for melanin biosynthesis in RPE melanosomes

(Fuhrmann, Zou, and Levine 2014, Strauss 2005, Chiba 2014, Sparrow, Hicks, and Hamel 2010). Microarray analysis of gene expression revealed that the knockdown of the eight candidate master TFs had somewhat different quantitative effects (Figure 2C), but there was a common set of ~1700 differentially expressed genes (FDR of 0.01 with absolute log2-fold change ≥ 1) (Figure 2D, Table S4), suggesting that RPE cells are similarly dependent on these factors for expression of this core set of genes. Examination of the down-regulated genes in this core set of genes showed significant enrichment of signature genes important for RPE function (Figures 2D and 2E). This RPE signature consisted of 154 highly expressed RPE genes previously identified by comparing the gene profiles of RPE cells to the Novartis expression database of 78 tissues (SymAtlas: http://wombat.gnf.org/index.html) (Strunnikova et al. 2010). In contrast, the up-regulated genes were associated with apoptotic cell death and cellular defense responses (Figures 2D and 2F). The morphological features of the cells were consistent with the induction of an apoptotic cell death program. These results indicate that the knockdown of the eight candidate master TFs caused a loss of the RPE cell expression program and subsequent induction of apoptosis. The similarity of the effects on gene expression observed with knockdown of these eight TFs suggests that they play similarly important roles in maintenance of the RPE gene expression program.

Studies of master TFs in embryonic stem cells and several differentiated cell types suggest that these factors share three common features (Lee and Young 2013, Whyte et al. 2013). These factors bind enhancers for a substantial

fraction of the genes that are actively transcribed, they bind clusters of enhancers (super-enhancers) at genes with prominent roles in cell-type specific biology, and they often bind the enhancers of their own genes as well as those of the other master TFs, thus forming a core circuitry of interconnected autoregulatory loops. To determine if the RPE candidate master TFs share these features, we identified RPE enhancers genome-wide and investigated the association of the RPE TFs with these enhancers (Figure 3A). Active enhancers were identified by using chromatin immunoprecipitation coupled to massively parallel sequencing (ChIP-Seq) with antibodies against the histone modification H3K27ac (Table S5), a nucleosomal modification that occurs at active enhancers (Creyghton et al. 2010, Rada-Iglesias et al. 2011). The results indicated that RPE cells have at least 17,679 sites with high confidence signal for histone H3K27ac (Figure 3A). We then carried out ChIP-Seq for the candidate master TFs and were able to obtain good quality data for five of the TFs (PAX6, LHX2, OTX2, MITF and ZNF92)(Figure 3A). The high confidence data revealed that these five candidate master TFs together occupied at least one third of the ~17,500 active enhancers (Figure 3B).

To determine whether the candidate master TFs bind super-enhancers at their own genes and those of other key cell identity genes, the ChIP-seq signal for H3K27ac was used to identify super-enhancers and their associated genes (Figure 3C, Table S6). The ChIP-seq data for the TFs was used to ascertain the pattern of TF binding to these super-enhancers (Figure 3D). The RPE super-enhancers occurred at many genes associated with RPE transcriptional control,

including the candidate master transcription factors SIX3, LHX2, OTX2 and FOXD1, and genes that feature prominently in RPE biology, including the retinal reductase gene DHRS3 (Figure 3C, Table S6). Examination of the super-enhancers revealed that different combinations of the five TFs occupied the various enhancer components of the super-enhancers (Figure 3D), as has been observed for master TFs at ESC super-enhancers (Whyte et al. 2013).

We next investigated whether the five candidate master TFs bind enhancers associated with their own genes as well as those associated with the other master TFs. The genome-wide binding data revealed that PAX6, LHX2 and OTX2 occupy active enhancers of genes encoding all five factors studied here, while MITF and ZNF92 occupied a subset of these enhancers (Figure 3E). Thus, the RPE master TF candidates form a core circuit with interconnected autoregulatory loops whose characteristics are similar to those previously described for other well-studied cells such as ESCs (Lee and Young 2013), hepatocytes (Odom et al. 2006), hematopoietic stem cells and erythroid cells (Novershtern et al. 2011) and T cell acute lymphoblastic leukemia cells (Sanda et al. 2012). A map of extended regulatory circuitry can be constructed for RPEs that includes genes that are both co-bound by all these regulators and dependent on their expression (Figure 3E; Table S7).

These results show that the RPE transcription factors studied here share key features with established master transcription factors, including binding to a large fraction of active enhancers, occupancy of super-enhancers at their own

genes and those of other key cell identity genes, and formation of core circuitry with interconnected autoregulatory loops.

Reprogramming of fibroblasts into RPE-like cells

Ectopic expression of master TFs can, for many cell types, reprogram gene expression programs and produce cells with functional states like those that normally express those master TFs (Vierbuchen and Wernig 2012, Buganim, Faddah, and Jaenisch 2013, Morris and Daley 2013, Sancho-Martinez, Baek, and Izpisua Belmonte 2012, Yamanaka 2012, Graf and Enver 2009). We therefore investigated whether the nine top scoring RPE master TF candidates can reprogram fibroblasts into an RPE-like state (Figure 4). Human foreskin fibroblasts (HFF) were transduced with an inducible doxycycline lentiviral cocktail with constructs for the nine TFs (Figure 4A). Colonies showing a "cobblestone"-like morphology characteristic of RPE cells were evident after two weeks of doxycycline induction. These colonies increased in size over two months in culture (Figure 4A). Independent cobblestone RPE-like colonies were manually picked and further expanded into six independent RPE-like cell lines. All six cell lines were found to contain the PAX6, OTX2, MITF, SIX3, GLIS3 and FOXD1 expression constructs (Figure 4B, Table S8) and to be able to maintain an RPE-like morphology in the presence of doxycycline for over 6 months (twelve passages). Two of the induced RPE-like cell lines, iRPE-1 and iRPE-2, were subjected to further analysis. Interestingly, these two iRPE lines were found to express all nine master TFs, suggesting the endogenous core circuitry was activated.

Initial analysis of the iRPE cell lines exhibited characteristic membrane expression of ZO-1 together with a "cobblestone" sheet morphology involving individual cells connected by tight junctions (Figure 4C). ZO-1 is a membrane-associated tight junction adaptor protein that links junctional membrane proteins to the cytoskeleton and signaling proteins. In RPE cells, these tight junctions have a fundamental role because they regulate paracellular diffusion across the blood-retinal barrier necessary for preventing substances from entering the retina (Harhaj and Antonetti 2004). The iRPE cells showed co-expression of CRALBP and RPE65 (Figure 4D), consistent with a functional visual cycle in these iRPE cells (Sparrow, Hicks, and Hamel 2010, Strauss 2005).

The iRPE-1 and iRPE-2 lines were subjected to gene expression analysis to determine if these cells produce the full RPE gene expression program. Principal component analysis (PCA) was carried out to compare the gene expression programs of the iRPE cells to those of 106 different cell types from Human Body Index collection, together with some additional well-studied cell types as positive and negative controls (Table S9). PCA revealed that the gene expression profiles of the two iRPE lines were as similar to RPE cells as iPSCs are to ESCs (Figure 4E). We focused further analysis on the genes that show differential expression between HFF and the RPEs. We found that expression data from the iRPE lines exhibited the gene expression signature found in normal RPE cells (Figure 4F).

iRPE function

RPE cells play crucial roles in the maintenance and function of retinal photoreceptors, including phagocytosis of shed outer segments of photoreceptors, transepithelial transport of nutrients and ions between the neural retina and the blood vessels, and secretion of growth factors and hormones. To test if iRPE cells can perform typical RPE functions, we cultured iRPE cells and RPE cells in transwells for 8 weeks to obtain RPE sheets. We then tested whether the iRPE cells were capable of phagocytosis of photoreceptor rod outer segments, able to form a barrier for ion transport, and capable of polarized hormone secretion (Figure 5).

Phagocytosis of photoreceptor rod outer segments (ROS) by RPE is essential for retinal function (Bok 1993). The essential role of RPE phagocytosis is highlighted by the rapid degeneration of photoreceptor neurons and subsequent blindness occurring in Royal College of Surgeons rats, which carry an autosomal recessive mutation that impairs RPE phagocytosis (Bok and Hall 1971). To test if iRPE cells can perform phagocytosis, we incubated mouse ROS with iRPE cells or HFF cells and tested for ROS incorporation using an antibody rhodopsin. Both iRPE cell lines stained positive for rhodopsin, indicating binding and incorporation of ROS into the RPE cells by phagocytosis (Figure 5A).

The RPE has structural properties of an ion transporting epithelium that controls transport of ions and water from the subretinal space, or apical side, to the blood vessels or basolateral side (Strauss 2005). Tight junctions between cells prevent ion and water movement between the apical and basolateral sides

of the cells. We evaluated this barrier function by measuring the transepithelial electrical resistance (TER), which provides a method to detect functional tight junctions (Stevenson et al. 1986). iRPE and RPE cells were cultured in transwells for 8 weeks prior to TER measurements. The mean TER was 275.6 ±17 $\Omega.cm^2$ and 232.2±10 $\Omega.cm^2$ for iRPE 1-2 clones, respectively, and 211.4±5 $\Omega.cm^2$, for RPE cells (Figure 5B). Thus, the iRPE cells were able to form an effective a barrier for ion transport and this was as effective as that observed for RPE cells.

The RPE produces and secretes a variety of growth factors and hormones to the apical and basolateral sides to maintain the structural properties of the retinal and blood vessels respectively (Ford et al. 2011). Vascular endothelial growth factor (VEGF) is released to the basolateral side preferentially and functions to prevent endothelial cell apoptosis in the blood vessels (Saint-Geniez et al. 2009). We cultured iRPE cells and RPE cells (Salero et al. 2012) in transwells and analyzed VEGF concentration secreted into the media from both apical and basolateral sides using ELISA. VEGF levels were 2,150 ± 190 and 2660 ± 63 pg/ml for the apical and basolateral sides respectively for iRPE-1, 1,731 ± 5 and 3050 ± 226 pg/ml for the apical and basolateral side respectively for iRPE-2 and 3,835 ± 190 and 5548 ± 691 for the apical and basolateral side respectively for RPE (Figure 5C), indicating a polarized secretion of VEGF in the iRPE lines that is similar to that produced by RPE cells.

We conclude that the iRPE cell lines are capable of three functions established for RPE cells: phagocytosis of photoreceptor rod outer segments, formation of a barrier for ion transport, and polarized growth factor secretion.

DISCUSSION

The retinal pigment epithelium provides vital support to photoreceptor cells and its dysfunction is associated with the onset and progression of age-related macular degeneration and other retinal dystrophies. We undertook a study of the master transcription factors of RPE cells to improve our understanding of the control of RPE gene expression and to explore whether these factors might facilitate generation of functional RPE-like cells from fibroblasts. RPE candidate master transcriptional regulators were identified using a novel computational method and these were used to guide exploration of the transcriptional regulatory circuitry of RPE cells, core features of which we describe here. The candidate master transcriptional regulators were also used to reprogram human fibroblasts into RPE-like cells (iRPEs). The iRPE cells share key features with RPEs derived from healthy individuals, including morphology, gene expression and functional attributes, and thus represent a step toward the goal of generating patient-matched RPE cells for treatment of macular degeneration.

The control of gene expression programs is apparently dominated by a small number of master transcription factors, but these have yet to be identified for most cell human types (Vierbuchen and Wernig 2012, Buganim, Faddah, and Jaenisch 2013, Morris and Daley 2013, Sancho-Martinez, Baek, and Izpisua

Belmonte 2012, Yamanaka 2012, Graf and Enver 2009). To identify candidate master TFs for the large population of human cell types, we devised a computational approach that exploits the observation that known master transcription factors are expressed at high levels in those cell types that have been well-studied. This approach examines the relative levels and cell-type-specificity of transcription factor expression in a large population of different cell types. With this method, we obtained an atlas of candidate master transcription factors for each of more than 100 cell types (Table S1, Table S2). This computational method is modular and scalable and thus can be adapted to predict master TFs for additional cell types for which expression data is not yet available.

The candidate master TFs for RPE cells were used to deduce key features the transcriptional regulatory circuitry of these cells. Knockdown experiments showed that these TFs play an important role in the expression of RPE signature genes identified previously (Strunnikova et al. 2010). These TFs occupied enhancers associated with a third of the actively transcribed RPE genes, bound super-enhancers at their own genes and those for additional genes with prominent roles in RPE cell identity, and formed a core regulatory circuitry with interconnected autoregulatory loops. These features are shared by master TFs of other well-studied cells (Lee and Young 2013, Novershtern et al. 2011, Sanda et al. 2012, Hnisz et al. 2013).

The RPE candidate master transcriptional regulators were used to reprogram human fibroblasts into iRPE cells that share key features with RPEs

derived from healthy individuals, including morphology, gene expression and functional attributes. The generation of iRPE cells is an important step toward the goal of more efficient generation of patient-matched RPE cells for treatment of macular degeneration and other retinal dystrophies. The generation of autologous transplantation strategies may have particular value for elderly patients, who are more susceptible to complications from the immunosuppressive treatments that often accompany other transplantation strategies. These iRPE cells require continuous activation of transgene expression to stably maintain their morphology over 6 months. Similar dependency on constitutive transgene activity has been observed for the transdifferentiated state in other cases (Sheng et al. 2012, Lujan et al. 2012, Buganim et al. 2012, Vierbuchen et al. 2010, Huang et al. 2011), and further optimization will be required to obtain transgene-independent lines for regenerative medicine applications. It is possible that other TFs that scored highly in the computational approach described above will facilitate full transgene-independent reprogramming.

For the vast majority of human cell types, the master transcription factors and the transcriptional programs they control is poorly understood. Furthermore, much of disease-associated sequence variation occurs in transcriptional regulatory regions (Farh et al. 2014, Maurano et al. 2012, Hnisz et al. 2013), but the transcriptional mechanisms that lead to disease pathology are understood in only a few instances. The atlas of candidate master TFs described here should therefore facilitate future exploration of the functions of key regulators of cell

identity, mapping of cellular regulatory circuitries and investigation of disease-associated mechanisms.

EXPERIMENTAL PROCEDURES

Identification of candidate master transcription factors

Briefly, an entropy-based measure of Jensen-Shannon divergence (Cabili et al. 2011) was adopted to identify candidate master transcription factors, based on the relative level and cell-type-specificity of expression of a given factor in one cell type compared to a background dataset of diverse human cell and tissue types. Expression datasets used are provided in Table S9. Additional details are provided in the Extended Experimental Procedures.

Cell culture

Human retinal pigment epithelial (RPE) cells used for ChIP-seq and knockdown experiments were purchased from ScienCell (ScienCell, cat. #6540). RPE cells were maintained in epithelial cell medium (EpiCM) (ScienCell, cat. #4101) supplemented with 2% fetal bovine serum (ScienCell, cat. #0010), 1x epithelial cell growth supplement (EpiCGS) (ScienCell, cat. #4152), and 1x penicillin/streptomycin solution (ScienCell, cat. #0503). Human foreskin fibroblasts (HFF) were purchased from GlobalStem (GlobalStem, cat. #GSC-3002) and maintained in DMEM (Life Technologies, cat. #11965-092) supplemented with 15% of Tet System Approved fetal bovine serum (Clontech, cat. #631101), 2mM L-Glutamine (Life Technologies, cat. #25030-081) and 100 U/ml penicillin-streptomycin (Life Technologies, cat. #15140-163).

Knockdown of candidate master transcription factors

shRNAmir lentiviral vectors were obtained from Thermo Scientific (Table S3). A non-targeting shRNAmir was used as a control. High-titer lentiviral particles for each plasmid were used to transduce RPE cells (ScienCell, cat. #6540). Twenty-four hours after infection, epithelial cell medium was replaced and selection with 1 µg/ml puromycin (Life Technologies, cat. #A1113803) was carried out. Puromycin-resistant cells were harvested for future analysis five days after transduction.

RNA Extraction, cDNA Preparation and Gene Expression Analysis

Total RNA from cultured cells was isolated using the RNeasy Mini Kit (Qiagen, cat. #74104), and cDNA was generated with SuperScript III First-Strand Synthesis System (Life technology, cat. #18080-051), following the manufacturer's suggested protocol. Quantitative real-time qPCR were carried out on the Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems) using gene-specific Taqman probes from Life Technologies (Table S10) and TaqMan Universal PCR Master Mix (Life Technologies, cat. #4364340), following the manufacturer's suggested protocol. For microarray analysis, total RNA was harvested and used for library preparation. For each transcription factor, total RNA was harvested from two different lines, each harboring a different shRNAmir construct. 100 ng of total RNA was used to prepare biotinylated cRNA (cRNA) using the 3' IVT Express Kit (Affymetrix, cat. #901228), following the manufacturer's suggested protocol. GeneChip Primeview Human Gene Expression Arrays (Affymetrix, cat. #901837) were hybridized and scanned

following the manufacturer's suggested protocols. Additional details are provided in the Extended Experimental Procedures.

ChIP-Seq and Analysis

Chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq) was performed as previously described (Lee, Johnstone, and Young 2006, Marson et al. 2008). Antibodies used for ChIP-seq are provided in Table S5. Additional details are provided in the Extended Experimental Procedures.

Construction of lentivirus-inducible vectors and ectopic expression experiments

The Lenti-X Tet-On Advanced Inducible Expression System (Clontech, cat. #632162) was used for ectopic expression experiments. For construction of lentiviral vectors, the inducible vector backbone (pLVX-Tight-Puro) was first modified to include an MluI site in the linker region for potential future cloning steps. Next, plasmids containing the full coding sequence of PAX6, OTX2, LHX2, MITF, SIX3, SOX9, GLIS3, FOXD1, or ZNF92 were obtained from Open Biosystems, Origene or the Dana Farber/Harvard Cancer Center DNA Resource Core (Table S11). Coding DNA sequences were amplified using oligos that also added small regions of DNA homologous to regions flanking the MluI site in the target vector (Table S11). Target vector was then cut with MluI and the amplified coding DNA sequences were cloned into the target vector via homologous recombination using the In-Fusion cloning system (Clontech, cat# 639646).

Expression plasmids were transformed and maintained in STBL4 cells (Life Technologies, cat# 11635-018).

Viral Preparation and Transduction of HFF

For ectopic expression experiments, HFF were first infected with pLVX-Tet-On Advanced, expressing rtTA Advanced. Cells were grown in 1 mg/ml Geneticin® Selective Antibiotic (Life Technologies, cat. #10131035) for two weeks to select for cells harboring the plasmid.

For virus preparation, replication-incompetent lentiviral particles were packaged in 293T cells in the presence of the envelope, pMD2, and packaging, psPAX, plasmids. Viral supernatants from cultures 36, 48, 60 and 72 hours post-transfection were filtered through a 0.45 μM filter. High-titer virus preparations for all nine transcription factors were then added to HFF in the presence of 5 μg/ml of polybrene (day 1). A second transduction with virus for all nine factors was performed the next day (day 2). After two days, transduced HFF were split and transferred to iRPE growth medium (see below)(day 3). The following day iRPE medium was supplemented with 2mg/ml doxycycline (Sigma Aldrich, cat. #D9891) (day 4). Medium was replaced every 3 days and fresh doxycycline added with every medium replacement.

iRPE growth conditions

iRPE lines were plated on Matrigel Basement Membrane Matrix-coated plates (BD, Cat. #CB-40234). iRPE cells were grown Minimum Essential Medium Eagle Alpha Modification (Sigma Aldrich, cat. #M4526) base medium containing 5% of Tet System Approved Fetal bovine serum (Clontech, cat. #631101), 1x N1

Medium Supplement (Sigma Aldrich, cat. #N6530), 1% Sodium Pyruvate (Life Technologies, cat. #11360070), 2mM L-Glutamine (Life Technologies, cat. #25030-081), 1x MEM Non-Essential Amino Acids (Life Technologies, cat. #11140), 1 mg/ml Geneticin® Selective Antibiotic (Life Technologies, cat #10131035), 100 U/ml penicillin-streptomycin (Life Technologies, cat. #15140-163) and THT (20 µg/L hydrocortisone (Sigma Aldrich, cat. #H6909), 250 mg/L taurine (Sigma Aldrich, cat. #T0625), and 0.013 µg/L triiodothyronine (Sigma Aldrich, cat. #T2877). Cells were incubated in a 37°C, 5% CO_2 humidified incubator.

Genotyping

To perform the genotyping of the iRPE lines, cells were lysed and genomic DNA was purified by treating samples with proteinase K, RNase A and phenol-chloroform extraction. DNA was amplified using GoTaq® Green Master Mix (Promega, cat. # M7122) using primers listed in Table S8. Primers were selected so one would hybridize in the coding region of the cDNA and the other would hybridize in the integrated viral sequence.

Immunostaining and Imaging

For immunostaining analysis, cells were grown in Corning® Transwell® polyester membrane cell culture inserts (Sigma Aldrich, cat. # CLS3460) for eight weeks in iRPE medium supplemented with 2 mg/ml doxycycline (Sigma Aldrich, cat. #D9891). Medium was replaced every three days. Cells plated in transwells were fixed in 4% paraformaldehyde for fifteen minutes on both apical and basal sides. Transwells inserts were then washed with 1x PBS three times for five

minutes. A 2mm biopsy punch of the transwell membrane was transferred to a glass slide. Slides were incubated in blocking/permeabilizing solution (1% BSA, 1% saponin and 5% normal goat serum in 1x PBS) for one hour at room temperature. Subsequently, primary antibodies were diluted in blocking/permeabilizing solution and incubated on the slides overnight at 4 °C. After three five-minute washes with 1x PBS, slides were incubated for one hour with appropriate Alexa secondary antibodies, diluted 1:500 in blocking/permeabilizing solution containing DAPI. Slides were then washed three times with 1x PBS and mounted with Prolong Gold Antifade Mountant (Life Technologies, cat. #P36930). Slides were left overnight at room temperature to solidify. Slides were visualized under a fluorescence microscope (Zeiss Axio Observer D1). Primary antibodies used for staining are listed in Table S5.

Phagocytosis Assay

Rod outer segments (ROS) were isolated following previously described protocols (Ryeom, Sparrow, and Silverstein 1996). Retinas were dissected immediately following sacrifice from 25 mice, ROS were isolated, and approximately 1.0 x 10^4 ROS were added to the supernatant of confluent cell cultures in transwells. The cells were then incubated for two hours at 37°C. Transwells were then washed 4-5 times with phosphate-buffered saline to remove all unbound ROS before fixation. Each transwell was fixed and immunostained for rhodopsin and dapi. Images were taken using fluorescence microscopy at a 40X magnification.

Transepithelial Electrical Resistance (TER)

iRPE cells were grown in Corning® Transwell® polyester membrane cell culture inserts (Sigma Aldrich, cat. # CLS3460) for eight weeks in iRPE medium supplemented with 2 mg/ml doxycycline (Sigma Aldrich, cat. #D9891). Medium was replaced every 3 days. Resistance was measured using the EVOM Epithelial Voltohmmeter (World Precision Instruments).

VEGF-A Release

iRPE cell and RPE cells (Salero et al., 2012) were grown in Corning® Transwell® polyester membrane cell culture inserts (Sigma Aldrich, cat. # CLS3460) for eight weeks in iRPE medium supplemented with 2 mg/ml doxycycline (Sigma Aldrich, cat. #D9891). Medium was replaced every three days with fresh doxycycline. Conditioned medium from apical and basal chambers of the same transwell insert was collected twenty-four hours following a complete medium change. VEGF-A protein secretion in conditioned medium was measured using a Human VEGF ELISA kit (Life Technologies, cat. #KHG0111), following the manufacturer's suggested protocol. Optical densities (450nm) were measured within two hours, using a microplate reader (Perkin Elmer 1420 Multilabel Counter). Data was analyzed using GraphPad Prism 6.

ACCESSION NUMBERS

Raw and processed sequencing and microarray data were deposited in GEO (Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo/), under

accession numbers GSE60024 and GSE64264 (reviewer link: http://www.ncbi.nlm.nih.gov/ geo/query/acc.cgi?token=ihklqeqivdydnmh&acc=GSE64264)

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures and 11 Supplemental tables.

Table S1. Catalog of candidate master transcription factors for cell types in the Human Body Index (GSE7307)

Table S2. Rank of top scoring candidate master transcription factors and additional reprogramming factors in a few well-studied cell types

Table S3. shRNAmir used in this study

Table S4. Gene expression changes in retinal pigment epithelial cells upon knockdown of candidate master transcription factors

Table S5. Antibodies used in this study

Table S6. RPE Super-enhancers and their associated genes in retinal pigment epithelial cells

Table S7. Genes bound by candidate master TFs and average expression changes upon single factor knockdown

Table S8. Genotyping primers

Table S9. Expression profiles used in the study

Table S10. Taqman probes used in this study

Table S11. Primers and CDNA used for construction of lentiviral vectors

AUTHOR CONTRIBUTIONS

A.C.D. contributed to the design of all experiments and performed knockdown, Chip-seq, and ectopic expression experiments. Z.P.F. contributed to the design of experiments, developed the method used to identify candidate master transcription factors and provided all bioinformatics-based analyses. K.J.W. performed the phagocytosis assay. M.A.C. assisted in the design of ectopic expression experiments and selection of iRPE colonies. J.S.S. performed staining for RPE markers and analysis of VEGF production. E.C. provided invaluable assistance in the maintenance of RPE and iRPE lines. C.C. performed the transepithelial resistance experiments. D.D. contributed to the design of experiments and computational methods and contributed to experiments. N.M.H. generated the lentiviral constructs used for ectopic expression experiments. R.J. and S.T. contributed to the conceptual development of the study. T.I.L, together with R.A.Y., initially conceived the study and contributed to the design of experiments and the conceptual development of the study. A.C.D., Z.P.F., T.I.L. and R.Y. wrote and edited the manuscript. M.A.C., R.J. and S.T. contributed critical comments on the manuscript.

ACKNOWLEDGEMENTS

We thank Tom Volkert, Jennifer Love and Sumeet Gupta at the Whitehead Genome Technologies Core for Solexa sequencing; Timothy Blenkinsop, Bluma Lesch, Alla Sigova, and Stephen H. Tsang for experimental assistance; Yossi

Bouganim, Maya Mitalipova, Frank Soldner, Denes Hnisz and members of the Young lab for helpful discussion; Garrett M. Frampton and Prathapan Thiru for critical discussion on the curation of expression datasets; and Johanna Goldmann and Jessica Reddy for critical comments on the manuscript. This work was supported by the National Institutes of Health grants HG002668 (R.A.Y.) and CA146445 (R.A.Y. and T.I.L.) and a grant from the Skolkovo Foundation (R.A.Y. and R.J.). The authors declare competing financial interests: R.J. is a cofounder of Fate Therapeutics and an adviser to Stemgent and R.A.Y. is a founder of Syros Pharmaceuticals.

FIGURES

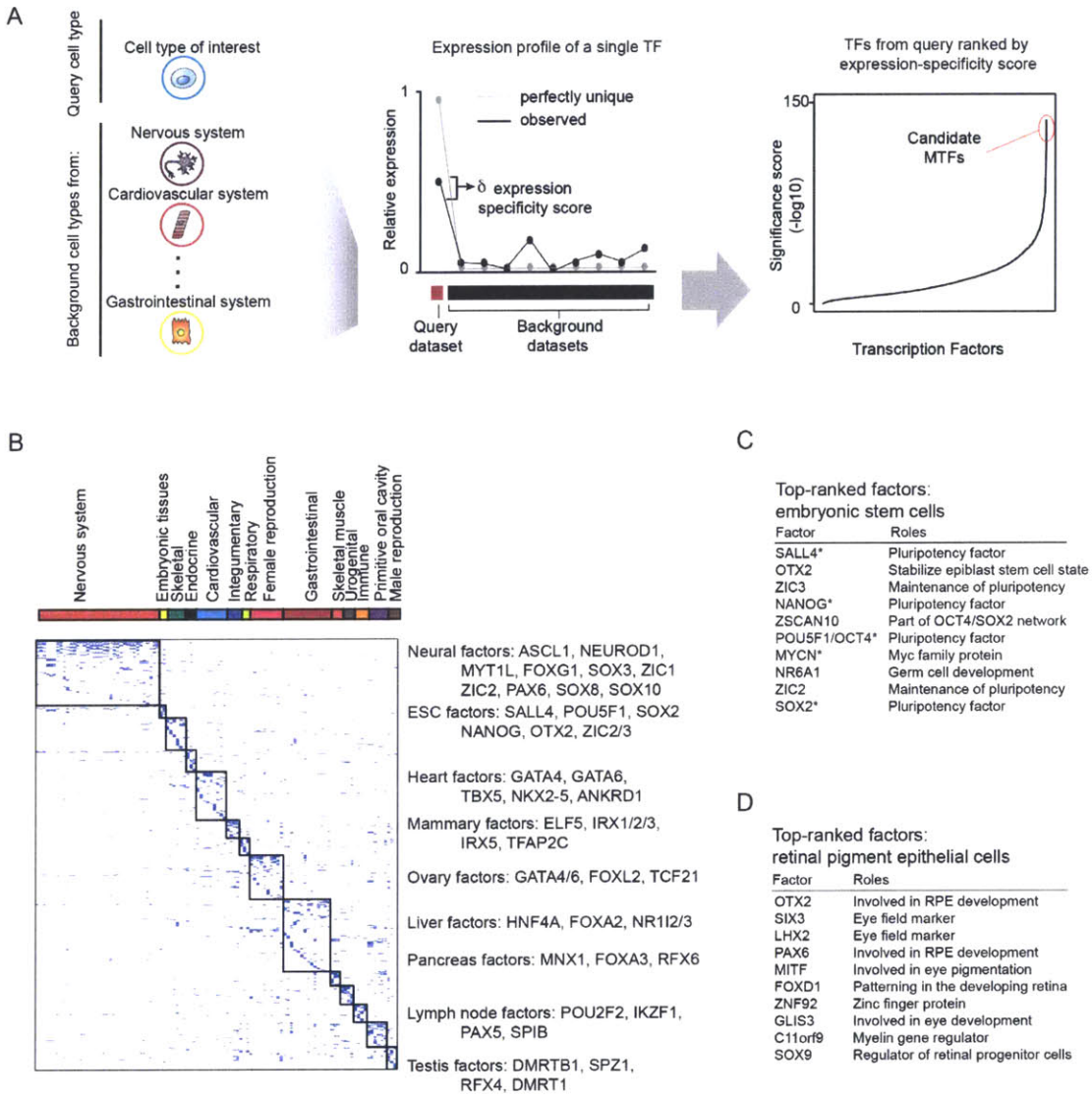


Figure 1. A general approach to identify candidate master transcription factors in human cells.

(A) Computational approach used to identify candidate master transcription factors in human cells.

Left panel: Collection of gene expression profiles of a query cell type and representative cell types from Human Body Index collection of expression data.

Middle panel: Expression profile of a single transcription factor across a query dataset and a range of background datasets. The idealized case of expression level of a transcription factor (grey) is compared to the observed data to calculate the expression-specificity score of the transcription factor.

Right panel: Plot depicting the distribution of significance scores of expression-specificity for all transcription factors. Factors are arranged on the x-axis in order of significance scores. Significance scores are indicated on the y-axis. The highest scoring transcription factors are considered the best candidate master transcription factors and highlighted in the red circle.

(B) Representation of the collection of candidate master transcription factors for 106 tissue and cell types. Tissue and cell types are arranged on the x-axis and clustered according to anatomical groups, represented by the colored bar at the top. Genes are arranged on the y-axis. Blue dashes represent candidate master transcription factors in a cell type. Clusters of candidate master transcription factors in cell types representing an anatomical group are boxed. Representative genes are listed on the side.

(C) List of top-scoring transcription factors in human ESCs ranked by expression specificity score. Asterisk indicates that the factor has been used in reprogramming experiments.

(D) List of top-scoring transcription factors in RPE cells ranked by expression specificity score.

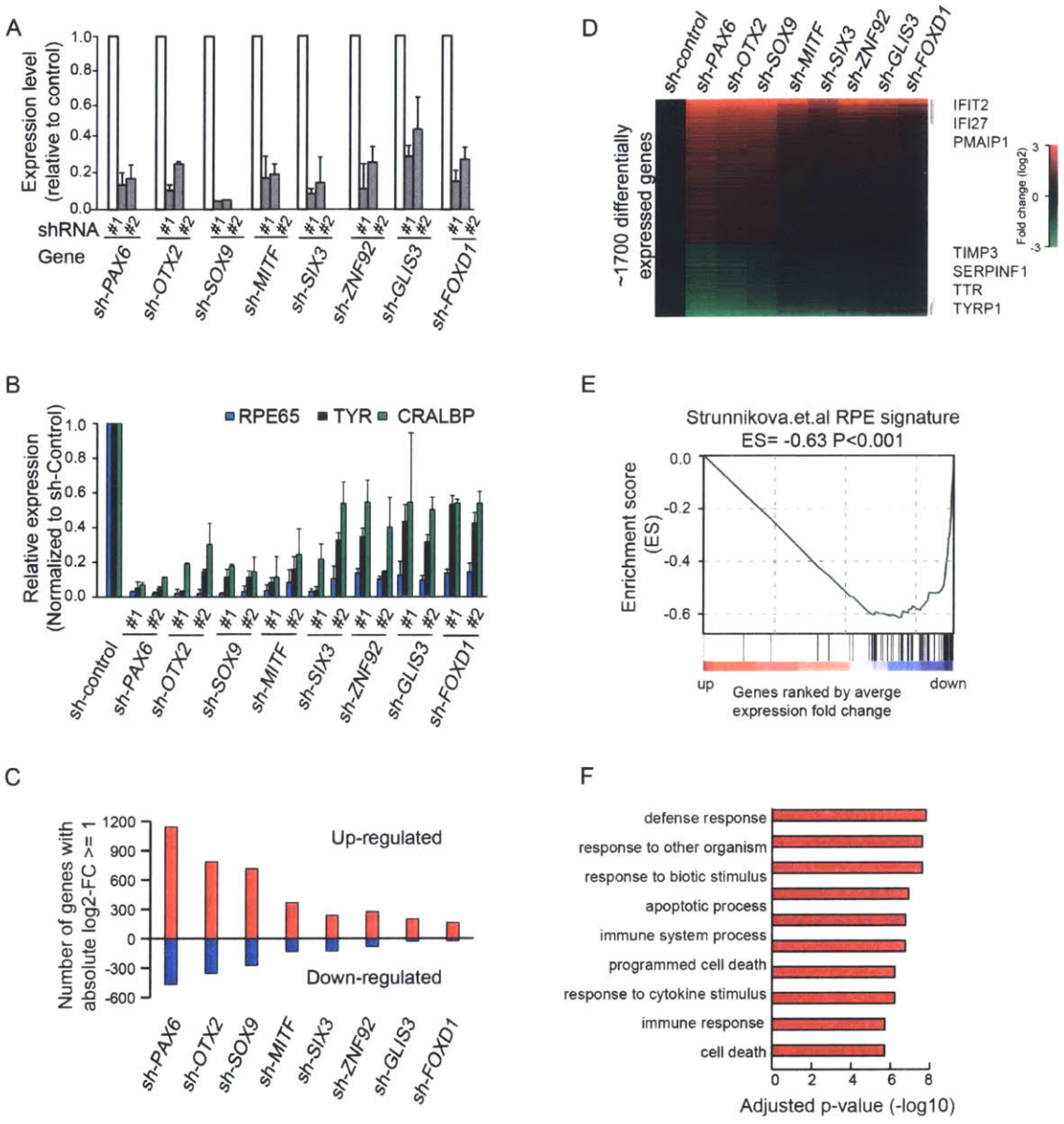


Figure 2. Maintenance of RPE identity identity depends on candidate master transcription factors.

(A) qPCR validation of knockdown efficiency at 5 days post-infection with shRNA lentiviruses. The percent knockdown for two independent shRNA lentiviral constructs (1 and 2) for each candidate master transcription factor is shown. Results are normalized to a non-targeting shRNA control. All error bars reflect s.d. (n=2).

B) RT-PCR expression analysis of the expression of transcripts of key RPE genes RPE65, TYR and CRALBP at 5 days post-infection with shRNA lentiviruses for candidate master TFs. Two independent shRNAs lentiviral constructs were used to knockdown each candidate master TF. Gene expression was normalized to GAPDH and calculated as a percent relative to non-targeting shRNA control ± SD (n=2).

(C) Bar plot showing the number of differentially expressed genes that have absolute log2-fold change ≥ 1 relative to the non-targeting shRNA control following the knockdown of each of the eight candidate master TFs.

(D) Global gene expression analysis of RPE cells at 5 days post-infection with shRNA lentiviruses for candidate TFs. The heatmap indicates the fold change (log2) of gene expression relative to the non-targeting shRNA control. Differentially expressed genes were combined and arranged in rows. The knockdown for each candidate master transcription factor or a non-targeting shRNA control are shown in columns. Knockdowns of candidate TFs cause reduced expression of key RPE genes including TIMP metallopeptidase inhibitor 3 (TIMP3), serpin peptidase inhibitor clade F member 1 (SERPINF1), transthyretin (TTR) and tyrosinase-related protein 1 (TYRP1) and increased

expression of apoptotic genes including interferon-induced protein with tetratricopeptide repeats 2 (IFIT2), interferon, alpha-inducible protein 27 (IFI27) and phorbol-12-myristate-13-acetate-induced protein 1 (PMAIP1).

(E) GSEA of differentially expressed RPE genes at 5 days post-infection with shRNA lentiviruses. The differentially expressed genes after the knockdown of each candidate master transcription factor were combined and pre-ranked by the average fold changes across experiments relative to a non-targeting shRNA control. An RPE-signature gene set (n = 152) from a previously published RPE transcriptome analysis (Strunnikova et al. 2010) was shown to be significantly down-regulated.

(F) Barplot showing the adjusted p-values (-log10) of the top10 enriched gene ontology terms for biological processes that are associated with the up-regulated genes after knockdown of candidate master transcription factors.

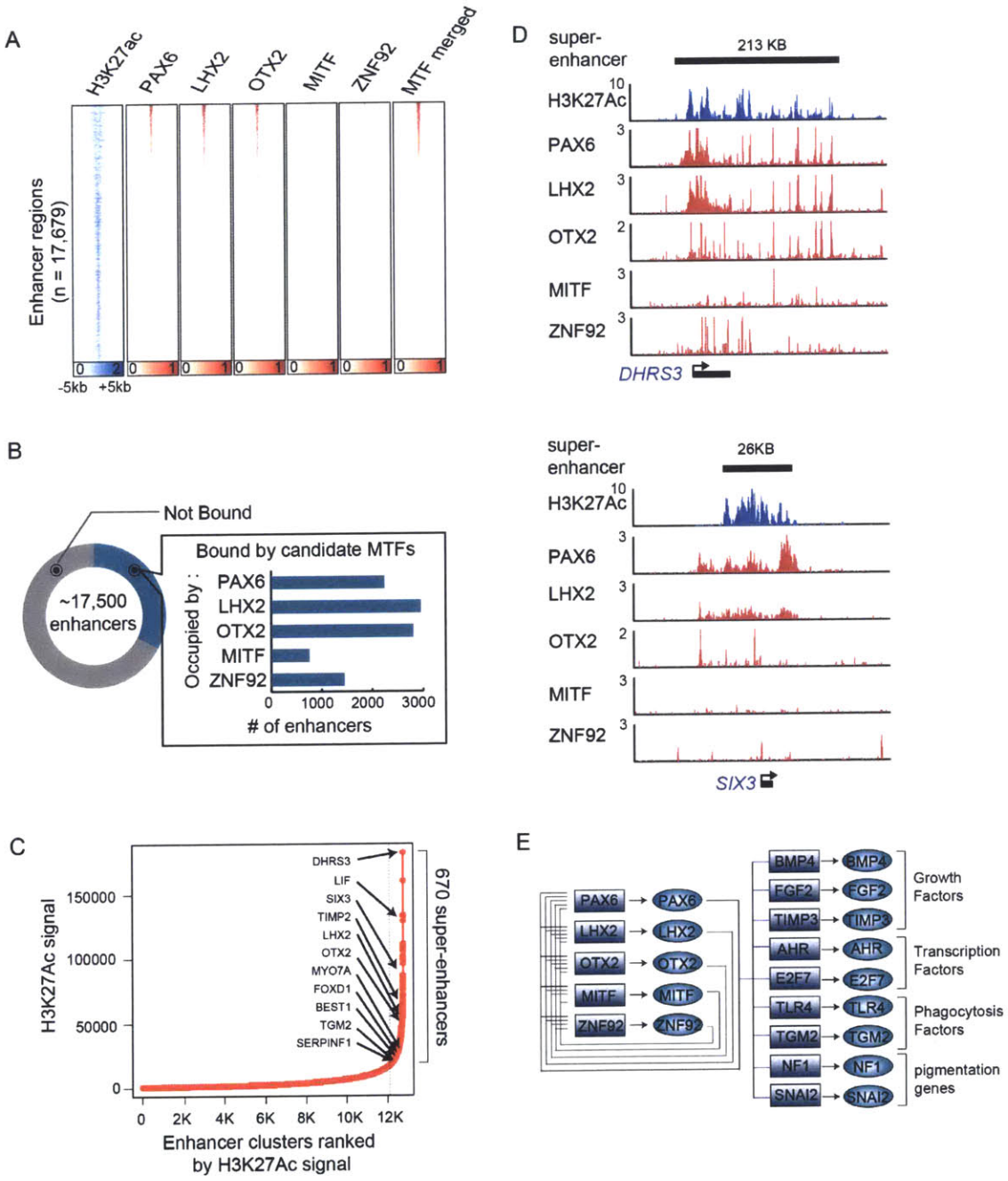


Figure 3. The transcriptional regulatory circuitry of human retinal pigment epithelial cells.

(A) Heat map showing the binding patterns for candidate master transcription factors at putative enhancer regions that show enrichment of H3K27AC (n= 17,679). ChIP-seq read density is shown for a 5-kb span, centered on the

putative enhancer regions. Color scale indicates ChIP-seq signal in units of rpm/bp.

(B) The overlap of the bound regions of PAX6, LHX2, OTX2, MITF, and ZNF92 with putative enhancer regions that show enrichment of H3K27AC (n= 17,679). Bar plot depicts the number of putative enhancer regions that are bound by each transcription factor.

(C) Distribution of H3K27ac ChIP-seq signal across the 12,750 enhancer clusters. ~17,500 active enhancers were stitched together into 12,750 enhancer clusters to identify super-enhancers (see Extended Experimental Procedures). Increasing background-subtracted H3K27ac ChIP-seq signal was used to rank the enhancer clusters. 670 super-enhancers containing exceptionally high amounts of H3K27ac were identified. Sample genes associated with RPE biology and their respective super-enhancers are highlighted.

(D) Tracks showing ChIP-seq enrichment of the active enhancer mark H3K27Ac at selected gene loci together with the signal for PAX6, LHX2, OTX2, MITF, and ZNF92. ChIP-seq signals are shown on the y-axis in units of reads per million mapped reads per base pair (rpm/bp). The location and size of the super-enhancer is shown at the top of the tracks and gene models are shown at the bottom.

(E) A model for the core transcriptional regulatory circuitry of RPE cells. Interconnected loops are formed by PAX6, LHX2, OTX2, MITF, and ZNF92. Genes are represented by rectangles and proteins are represented by ovals.

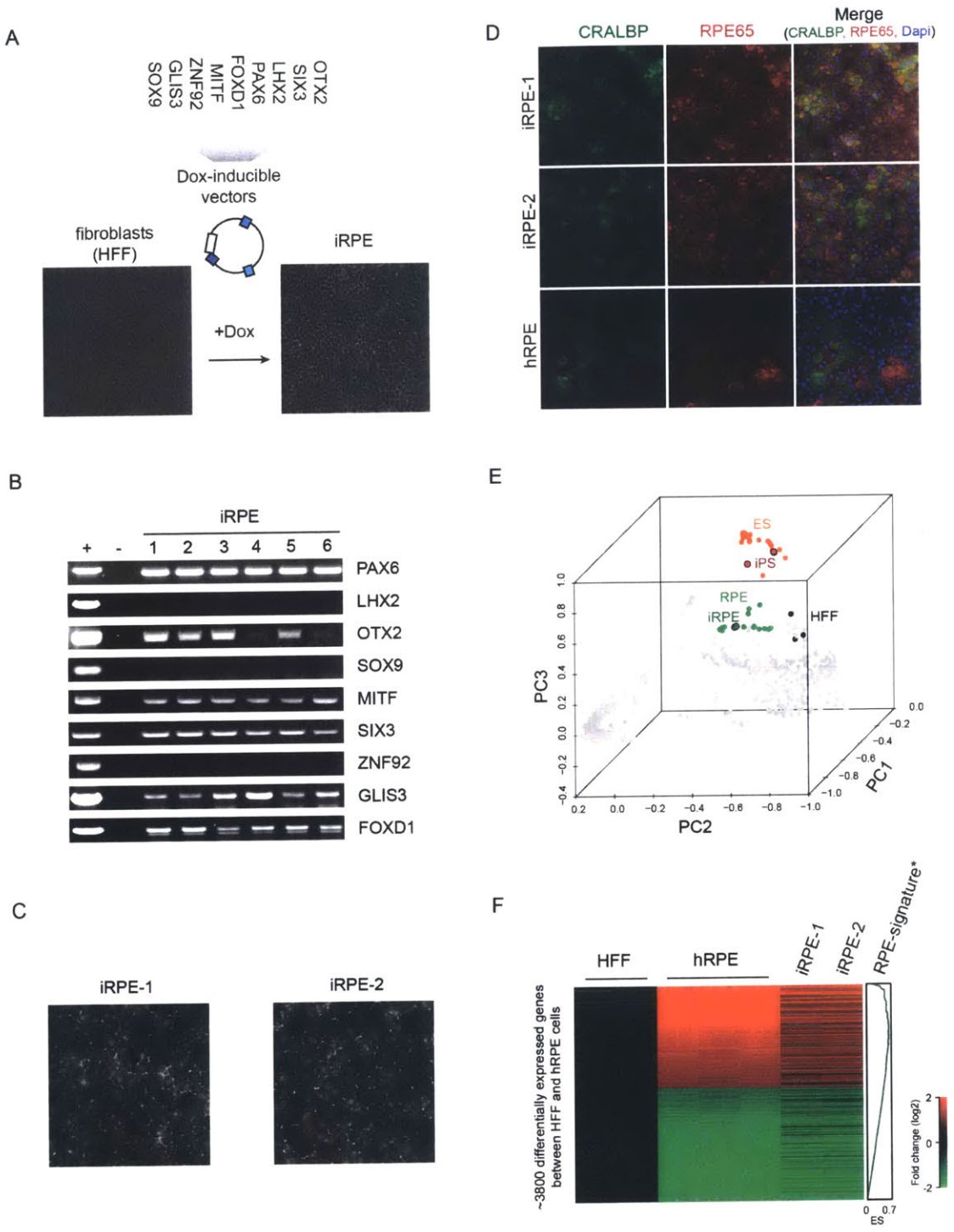


Figure 4. Ectopic expression of RPE candidate master transcription factors is sufficient to drive the morphology and gene expression program of fibroblasts towards an RPE-like state.

(A) Schematic outlining the ectopic expression of candidate master transcription factors in human neonatal foreskin fibroblasts (HFF). Lentiviral constructs were induced to express candidate master transcription factors with doxycycline (Dox).

(B) PCR analysis of transgene integration for iRPE lines. DNA of the constructs used to make the lentivirus is shown as a positive control.

(C) Immunostaining imaging of iRPE-1 and iRPE-2 cells in the presence of Dox. Cells were immunostained with Z0-1.

(D) Immunostaining imaging of RPE, iRPE-1 and iRPE-2 cells in the presence of Dox. Cells were immunostained with retinal pigment epithelial cells markers CRALBP, RPE65 and dapi.

(E) Principle component analysis (PCA) comparing the gene expression profiles of iRPE cells to gene expression profiles of over 100 other cell types. The expression profiles of HFF (Parker et al.), iRPE cells (green), RPE (light green), iPS (Reddy et al.) and ES (orange red), 106 additional cell types (grey) were shown in the PCA plot.

(D) Global gene expression analysis of retinal pigment epithelial cells and fibroblasts. Differentially expressed genes of HFF and RPE are arranged along the rows. Different expression profiles are shown in columns. The heat map indicates the fold change (log2) of gene expression relative to the HFF control. GSEA enrichment score of a previously published RPE signature gene set (Strunnikova et al. 2010) is shown in the rightmost column.

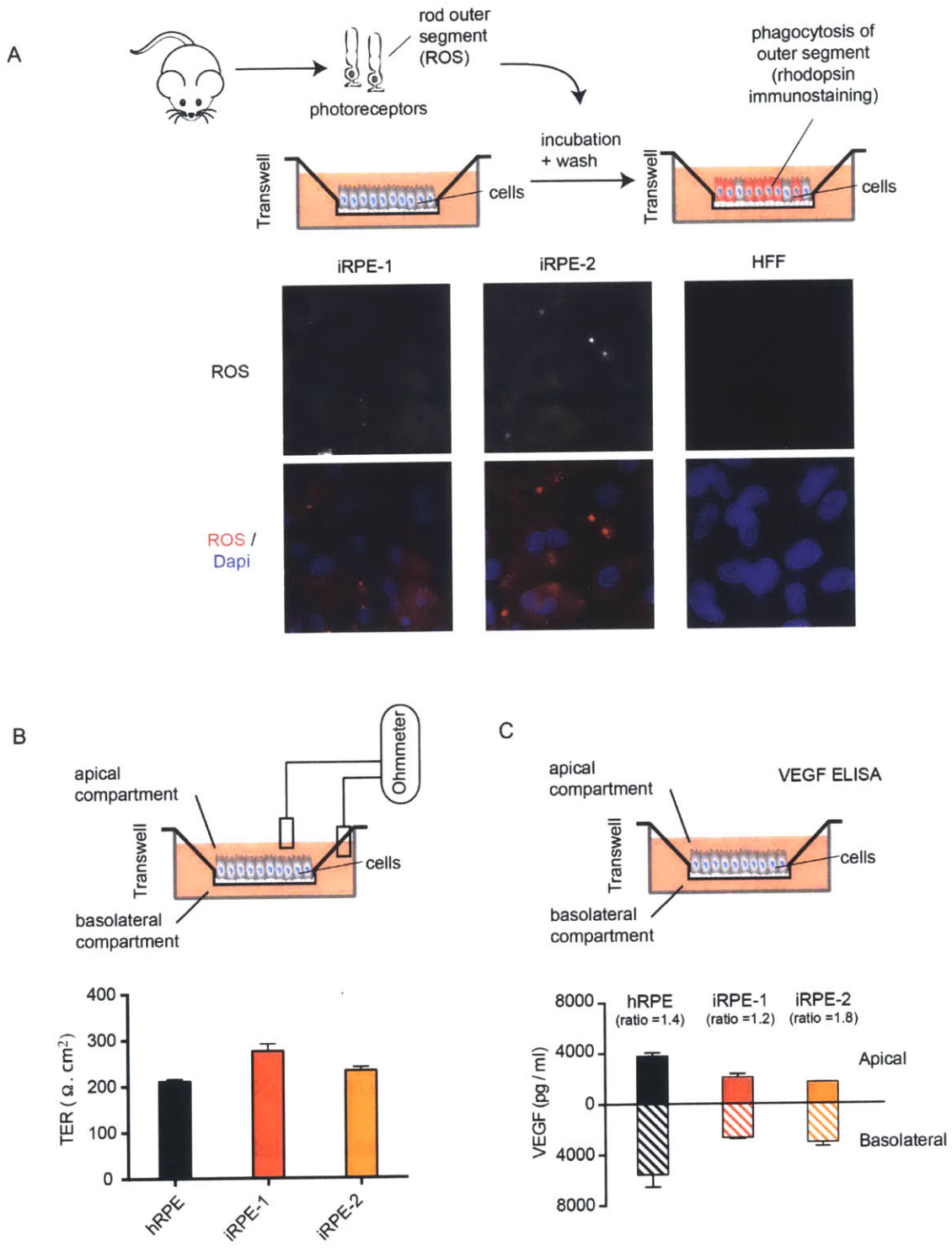


Figure 5. RPE-like cells have functional characteristics.

(A) iRPE cells demonstrate phagocytosis of photoreceptor outer segments. Photoreceptor outer segments were dissected from mice and incubated with iRPE and HFF cells for two hours. Immunostaining for rhodopsin and DAPI are shown.

(B) Trans-epithelial resistance (TER) for iRPE-1, iRPE-2 and RPE (Salero et al. 2012) . TER for iRPE-1, iRPE-2 and RPE cells (mean + SD) is 275.6 ±17 Ω.cm2, 232.2±10 Ω.cm2 and 211.4±5 Ω.cm2 respectively.

(C) iRPE cells secrete vascular endothelial growth factor (VEGF) in a polarized manner. Conditioned media were collected from the apical and basolateral transwell compartments and analyzed for VEGF with enzyme-linked immunosorbent assay (ELISA). A preferential secretion of VEGF toward the basolateral side was observed for both iRPE-1, iRPE-1 and RPE (Salero et al. 2012). The ratio of VEGF release between the basolateral and apical sides is shown above each bar.

REFERENCE

Avilion, A. A., S. K. Nicolis, L. H. Pevny, L. Perez, N. Vivian, and R. Lovell-Badge. 2003. "Multipotent cell lineages in early mouse development depend on SOX2 function." *Genes Dev* 17 (1):126-40. doi: 10.1101/gad.224503.

Bharti, K., M. Gasper, J. X. Ou, M. Brucato, K. Clore-Gronenborn, J. Pickel, and H. Arnheiter. 2012. "A Regulatory Loop Involving PAX6, MITF, and WNT Signaling Controls Retinal Pigment Epithelium Development." *Plos Genetics* 8 (7). doi: Artn E1002757 Doi 10.1371/Journal.Pgen.1002757.

Binder, S., B. V. Stanzel, I. Krebs, and C. Glittenberg. 2007. "Transplantation of the RPE in AMD." *Prog Retin Eye Res* 26 (5):516-54. doi: 10.1016/j.preteyeres.2007.02.002.

Bok, D. 1993. "The retinal pigment epithelium: a versatile partner in vision." *J Cell Sci Suppl* 17:189-95.

Bok, D., and M. O. Hall. 1971. "The role of the pigment epithelium in the etiology of inherited retinal dystrophy in the rat." *J Cell Biol* 49 (3):664-82.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. 2005. "Core transcriptional regulatory circuitry in human embryonic stem cells." *Cell* 122 (6):947-56. doi: S0092-8674(05)00825-1 [pii] 10.1016/j.cell.2005.08.020.

Buganim, Y., D. A. Faddah, and R. Jaenisch. 2013. "Mechanisms and models of somatic cell reprogramming." *Nat Rev Genet* 14 (6):427-39. doi: 10.1038/nrg3473.

Buganim, Y., E. Itskovich, Y. C. Hu, A. W. Cheng, K. Ganz, S. Sarkar, D. Fu, G. G. Welstead, D. C. Page, and R. Jaenisch. 2012. "Direct reprogramming of fibroblasts into embryonic Sertoli-like cells by defined factors." *Cell Stem Cell* 11 (3):373-86. doi: S1934-5909(12)00477-8 [pii] 10.1016/j.stem.2012.07.019.

Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." *Genes Dev* 25 (18):1915-27. doi: 10.1101/gad.17446611.

Chambers, I., D. Colby, M. Robertson, J. Nichols, S. Lee, S. Tweedie, and A. Smith. 2003. "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." *Cell* 113 (5):643-55. doi: S0092867403003921 [pii].

Chiba, C. 2014. "The retinal pigment epithelium: an important player of retinal disorders and regeneration." *Exp Eye Res* 123:107-14. doi: 10.1016/j.exer.2013.07.009.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 2010. "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50):21931-21936. doi: Doi 10.1073/Pnas.1016071107.

Cyranoski, D. 2013. "Stem cells cruise to clinic." *Nature* 494 (7438):413. doi: 494413a [pii] 10.1038/494413a.

Cyranoski, D. 2014. "Stem-cell method faces fresh questions." *Nature* 507 (7492):283. doi: 10.1038/507283a.

da Cruz, L., F. K. Chen, A. Ahmado, J. Greenwood, and P. Coffey. 2007. "RPE transplantation and its role in retinal disease." *Prog Retin Eye Res* 26 (6):598-635. doi: 10.1016/j.preteyeres.2007.07.001.

Farh, K. K., A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. 2014. "Genetic and epigenetic fine mapping of causal autoimmune disease variants." *Nature*. doi: 10.1038/nature13835.

Ford, K. M., M. Saint-Geniez, T. Walshe, A. Zahr, and P. A. D'Amore. 2011. "Expression and role of VEGF in the adult retinal pigment epithelium." *Invest Ophthalmol Vis Sci* 52 (13):9478-87. doi: 10.1167/iovs.11-8353.

Fuhrmann, S., C. Zou, and E. M. Levine. 2014. "Retinal pigment epithelium development, plasticity, and tissue homeostasis." *Exp Eye Res* 123:141-50. doi: 10.1016/j.exer.2013.09.003.

Graf, T., and T. Enver. 2009. "Forcing cells to change lineages." *Nature* 462 (7273):587-94. doi: nature08533 [pii] 10.1038/nature08533.

Harhaj, N. S., and D. A. Antonetti. 2004. "Regulation of tight junctions and loss of barrier function in pathophysiology." *Int J Biochem Cell Biol* 36 (7):1206-37. doi: 10.1016/j.biocel.2003.08.007.

Henriques, T., D. A. Gilchrist, S. Nechaev, M. Bern, G. W. Muse, A. Burkholder, D. C. Fargo, and K. Adelman. 2013. "Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals." *Mol Cell* 52 (4):517-28. doi: 10.1016/j.molcel.2013.10.001.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. 2013. "Super-enhancers in the control of cell identity and disease." *Cell* 155 (4):934-47. doi: 10.1016/j.cell.2013.09.053.

Huang, P., Z. He, S. Ji, H. Sun, D. Xiang, C. Liu, Y. Hu, X. Wang, and L. Hui. 2011. "Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors." *Nature* 475 (7356):386-9. doi: 10.1038/nature10116.

Ivanova, N., R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I. R. Lemischka. 2006. "Dissecting self-renewal in stem cells with RNA interference." *Nature* 442 (7102):533-8. doi: nature04915 [pii] 10.1038/nature04915.

Iwafuchi-Doi, M., and K. S. Zaret. 2014. "Pioneer transcription factors in cell reprogramming." *Genes Dev* 28 (24):2679-2692. doi: 10.1101/gad.253443.114.

Kamao, H., M. Mandai, S. Okamoto, N. Sakai, A. Suga, S. Sugita, J. Kiryu, and M. Takahashi. 2014. "Characterization of human induced pluripotent stem cell-derived retinal pigment epithelium cell sheets aiming for clinical application." *Stem Cell Reports* 2 (2):205-18. doi: 10.1016/j.stemcr.2013.12.007.

Kim, J., J. Chu, X. Shen, J. Wang, and S. H. Orkin. 2008. "An extended transcriptional network for pluripotency of embryonic stem cells." *Cell* 132 (6):1049-61. doi: S0092-8674(08)00328-0 [pii] 10.1016/j.cell.2008.02.039.

Lee, T. I., S. E. Johnstone, and R. A. Young. 2006. "Chromatin immunoprecipitation and microarray-based analysis of protein location." *Nat Protoc* 1 (2):729-48. doi: nprot.2006.98 [pii] 10.1038/nprot.2006.98.

Lee, T. I., and R. A. Young. 2013. "Transcriptional regulation and its misregulation in disease." *Cell* 152 (6):1237-51. doi: 10.1016/j.cell.2013.02.014.

Lim, L. S., P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong. 2012. "Age-related macular degeneration." *Lancet* 379 (9827):1728-38. doi: S0140-6736(12)60282-7 [pii] 10.1016/S0140-6736(12)60282-7.

Lujan, E., S. Chanda, H. Ahlenius, T. C. Sudhof, and M. Wernig. 2012. "Direct conversion of mouse fibroblasts to self-renewing, tripotent neural precursor cells." *Proc Natl Acad Sci U S A* 109 (7):2527-32. doi: 10.1073/pnas.1121003109.

Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Young. 2008. "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." *Cell* 134 (3):521-33. doi: S0092-8674(08)00938-0 [pii] 10.1016/j.cell.2008.07.020.

Martinez-Morales, J. R., V. Dolez, I. Rodrigo, R. Zaccarini, L. Leconte, P. Bovolenta, and S. Saule. 2003. "OTX2 activates the molecular network underlying retina pigment epithelium differentiation." *Journal of Biological Chemistry* 278 (24):21721-21731. doi: Doi 10.1074/Jbc.M301708200.

Masuda, T., and N. Esumi. 2010. "SOX9, through Interaction with Microphthalmia-associated Transcription Factor (MITF) and OTX2, Regulates BEST1 Expression in the Retinal Pigment Epithelium." *Journal of Biological Chemistry* 285 (35):26933-26944. doi: Doi 10.1074/Jbc.M110.130294.

Matsuo, I., S. Kuratani, C. Kimura, N. Takeda, and S. Aizawa. 1995. "Mouse Otx2 Functions in the Formation and Patterning of Rostral Head." *Genes & Development* 9 (21):2646-2658. doi: Doi 10.1101/Gad.9.21.2646.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Z. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science* 337 (6099):1190-1195. doi: Doi 10.1126/Science.1222794.

Morris, S. A., and G. Q. Daley. 2013. "A blueprint for engineering cell fate: current technologies to reprogram cell identity." *Cell Res* 23 (1):33-48. doi: 10.1038/cr.2013.1.

Novershtern, N., A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, G. M. Frampton, A. C. Drake, I. Leskov, B. Nilsson, F. Preffer, D. Dombkowski, J. W. Evans, T. Liefeld, J. S. Smutko, J. Chen, N. Friedman, R. A. Young, T. R. Golub, A. Regev, and B. L. Ebert. 2011. "Densely interconnected transcriptional circuits control cell states in human hematopoiesis." *Cell* 144 (2):296-309. doi: S0092-8674(11)00005-5 [pii] 10.1016/j.cell.2011.01.004.

Odom, D. T., R. D. Dowell, E. S. Jacobsen, L. Nekludova, P. A. Rolfe, T. W. Danford, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. 2006. "Core transcriptional regulatory circuitry in human hepatocytes." *Mol Syst Biol* 2:2006 0017. doi: msb4100059 [pii] 10.1038/msb4100059.

Parker, S. C., M. L. Stitzel, D. L. Taylor, J. M. Orozco, M. R. Erdos, J. A. Akiyama, K. L. van Bueren, P. S. Chines, N. Narisu, Nisc Comparative Sequencing Program, B. L. Black, A. Visel, L. A. Pennacchio, F. S. Collins, Authors National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, and Nisc Comparative Sequencing Program Authors. 2013. "Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants." *Proc Natl Acad Sci U S A* 110 (44):17921-6. doi: 10.1073/pnas.1317023110.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. "A unique chromatin signature uncovers early developmental enhancers in humans." *Nature* 470 (7333):279-283. doi: Doi 10.1038/Nature09692.

Reddy, T. E., J. Gertz, F. Pauli, K. S. Kucera, K. E. Varley, K. M. Newberry, G. K. Marinov, A. Mortazavi, B. A. Williams, L. Y. Song, G. E. Crawford, B. Wold, H. F. Willard, and R. M. Myers. 2012. "Effects of sequence variation on differential allelic transcription factor occupancy and gene expression." *Genome Research* 22 (5):860-869. doi: Doi 10.1101/Gr.131201.111.

Rivera, C. M., and B. Ren. 2013. "Mapping human epigenomes." *Cell* 155 (1):39-55. doi: 10.1016/j.cell.2013.09.011.

Ryeom, S. W., J. R. Sparrow, and R. L. Silverstein. 1996. "CD36 participates in the phagocytosis of rod outer segments by retinal pigment epithelium." *J Cell Sci* 109 (Pt 2):387-95.

Saint-Geniez, M., T. Kurihara, E. Sekiyama, A. E. Maldonado, and P. A. D'Amore. 2009. "An essential role for RPE-derived soluble VEGF in the maintenance of the choriocapillaris." *Proc Natl Acad Sci U S A* 106 (44):18751-6. doi: 10.1073/pnas.0905010106.

Salero, E., T. A. Blenkinsop, B. Corneo, A. Harris, D. Rabin, J. H. Stern, and S. Temple. 2012. "Adult human RPE can be activated into a multipotent stem cell that produces mesenchymal derivatives." *Cell Stem Cell* 10 (1):88-95. doi: 10.1016/j.stem.2011.11.018.

Sancho-Martinez, I., S. H. Baek, and J. C. Izpisua Belmonte. 2012. "Lineage conversion methodologies meet the reprogramming toolbox." *Nat Cell Biol* 14 (9):892-9. doi: 10.1038/ncb2567.

Sanda, T., L. N. Lawton, M. I. Barrasa, Z. P. Fan, H. Kohlhammer, A. Gutierrez, W. Ma, J. Tatarek, Y. Ahn, M. A. Kelliher, C. H. Jamieson, L. M. Staudt, R. A. Young, and A. T. Look. 2012. "Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia." *Cancer Cell* 22 (2):209-21. doi: S1535-6108(12)00256-5 [pii] 10.1016/j.ccr.2012.06.007.

Schwartz, S. D., J. P. Hubschman, G. Heilwell, V. Franco-Cardenas, C. K. Pan, R. M. Ostrick, E. Mickunas, R. Gay, I. Klimanskaya, and R. Lanza. 2012. "Embryonic stem cell trials for macular degeneration: a preliminary report." *Lancet* 379 (9817):713-20. doi: S0140-6736(12)60028-2 [pii] 10.1016/S0140-6736(12)60028-2.

Schwartz, S. D., C. D. Regillo, B. L. Lam, D. Eliott, P. J. Rosenfeld, N. Z. Gregori, J. P. Hubschman, J. L. Davis, G. Heilwell, M. Spirn, J. Maguire, R. Gay, J. Bateman, R. M. Ostrick, D. Morris, M. Vincent, E. Anglade, L. V. Del Priore, and R. Lanza. 2014. "Human embryonic stem cell-derived retinal pigment epithelium in patients with age-related macular degeneration and Stargardt's macular dystrophy: follow-up of two open-label phase 1/2 studies." *Lancet*. doi: 10.1016/S0140-6736(14)61376-3.

Sheng, C., Q. Zheng, J. Wu, Z. Xu, L. Wang, W. Li, H. Zhang, X. Y. Zhao, L. Liu, Z. Wang, C. Guo, H. J. Wu, Z. Liu, L. Wang, S. He, X. J. Wang, Z. Chen, and Q.

Zhou. 2012. "Direct reprogramming of Sertoli cells into multipotent neural stem cells by defined factors." *Cell Res* 22 (1):208-18. doi: 10.1038/cr.2011.175.

Soufi, A., G. Donahue, and K. S. Zaret. 2012. "Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome." *Cell* 151 (5):994-1004. doi: 10.1016/j.cell.2012.09.045.

Sparrow, J. R., D. Hicks, and C. P. Hamel. 2010. "The retinal pigment epithelium in health and disease." *Curr Mol Med* 10 (9):802-23. doi: CMM # 78 [pii].

Stergachis, A. B., S. Neph, A. Reynolds, R. Humbert, B. Miller, S. L. Paige, B. Vernot, J. B. Cheng, R. E. Thurman, R. Sandstrom, E. Haugen, S. Heimfeld, C. E. Murry, J. M. Akey, and J. A. Stamatoyannopoulos. 2013. "Developmental fate and cellular maturity encoded in human regulatory DNA landscapes." *Cell* 154 (4):888-903. doi: 10.1016/j.cell.2013.07.020.

Stevenson, B. R., J. D. Siliciano, M. S. Mooseker, and D. A. Goodenough. 1986. "Identification of ZO-1: a high molecular weight polypeptide associated with the tight junction (zonula occludens) in a variety of epithelia." *J Cell Biol* 103 (3):755-66.

Strauss, O. 2005. "The retinal pigment epithelium in visual function." *Physiological Reviews* 85 (3):845-881. doi: Doi 10.1152/Physrev.00021.2004.

Strunnikova, N. V., A. Maminishkis, J. J. Barb, F. Wang, C. Zhi, Y. Sergeev, W. Chen, A. O. Edwards, D. Stambolian, G. Abecasis, A. Swaroop, P. J. Munson, and S. S. Miller. 2010. "Transcriptome analysis and molecular signature of human retinal pigment epithelium." *Hum Mol Genet* 19 (12):2468-86. doi: ddq129 [pii] 10.1093/hmg/ddq129.

Vierbuchen, T., A. Ostermeier, Z. P. Pang, Y. Kokubu, T. C. Sudhof, and M. Wernig. 2010. "Direct conversion of fibroblasts to functional neurons by defined factors." *Nature* 463 (7284):1035-41. doi: nature08797 [pii] 10.1038/nature08797.

Vierbuchen, T., and M. Wernig. 2012. "Molecular roadblocks for cellular reprogramming." *Mol Cell* 47 (6):827-38. doi: 10.1016/j.molcel.2012.09.008.

Wang, Z. X., J. L. Kueh, C. H. Teh, M. Rossbach, L. Lim, P. Li, K. Y. Wong, T. Lufkin, P. Robson, and L. W. Stanton. 2007. "Zfp206 is a transcription factor that controls pluripotency of embryonic stem cells." *Stem Cells* 25 (9):2173-82. doi: 2007-0085 [pii] 10.1634/stemcells.2007-0085.

Wang, Z. X., C. H. Teh, J. L. Kueh, T. Lufkin, P. Robson, and L. W. Stanton. 2007. "Oct4 and Sox2 directly regulate expression of another pluripotency transcription factor, Zfp206, in embryonic stem cells." *J Biol Chem* 282 (17):12822-30. doi: M611814200 [pii] 10.1074/jbc.M611814200.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. 2013. "Master transcription factors and

mediator establish super-enhancers at key cell identity genes." *Cell* 153 (2):307-19. doi: S0092-8674(13)00392-9 [pii] 10.1016/j.cell.2013.03.035.

Xie, W., and B. Ren. 2013. "Developmental biology. Enhancing pluripotency and lineage specification." *Science* 341 (6143):245-7. doi: 10.1126/science.1236254.

Yamanaka, S. 2012. "Induced pluripotent stem cells: past, present, and future." *Cell Stem Cell* 10 (6):678-84. doi: S1934-5909(12)00237-8 [pii] 10.1016/j.stem.2012.05.005.

Young, R. A. 2011. "Control of the embryonic stem cell state." *Cell* 144 (6):940-54. doi: 10.1016/j.cell.2011.01.032.

Zhang, K., G. H. Liu, F. Yi, N. Montserrat, T. Hishida, C. R. Esteban, and J. C. Izpisua Belmonte. 2014. "Direct conversion of human fibroblasts into retinal pigment epithelium-like cells by defined factors." *Protein Cell* 5 (1):48-58. doi: 10.1007/s13238-013-0011-2.

Chapter 3

Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes

Jill M. Dowen[1*], Zi Peng Fan[1,2*], Denes Hnisz[1*], Gang Ren[3,4*], Brian J. Abraham[1], Lyndon N. Zhang[1,5], Abraham S. Weintraub[1,5], Jurian Schujiers[1], Tong Ihn Lee[1], Keji Zhao[3], Richard A. Young[1,5]

[1] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA, [2] Computational and Systems Biology Program, [3] Systems Biology Center, NHLBI, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892, USA, [4] College of Animal Science and Technology, Northwest A&F University, Xi'An, P. R. China, [5] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

* These authors contributed equally

This chapter originally appeared in Cell 2014 vol. 159 (2) pp. 374-387. Corresponding Supplementary Material is appended. Supplementary tables and data are available online at http://dx.doi.org/10.1016/j.cell.2014.09.030.

Personal Contribution to the Project

This work was a close collaboration between Jill M. Dowen, Denes Hnisz and myself. I performed the majority of computational analyses, with assistance from Brian J. Abraham, and Lyndon N. Zhang. Jill M. Dowen, Denes Hnisz, Gang Ren, and Abraham S. Weintraub performed the experiments. The manuscript was written by Jill M. Dowen, Denes Hnisz, Richard A. Young, and myself.

SUMMARY

The pluripotent state of embryonic stem cells (ESCs) is produced by active transcription of genes that control cell identity and repression of genes encoding lineage-specifying developmental regulators. Here we use ESC cohesin ChIA-PET data to identify the local chromosomal structures at both active and repressed genes across the genome. The results produce a map of enhancer-promoter interactions and reveal that super-enhancer driven genes generally occur within chromosome structures that are formed by the looping of two interacting CTCF sites co-occupied by cohesin. These looped structures form insulated neighborhoods whose integrity is important for proper expression of local genes. We also find that repressed genes encoding lineage-specifying developmental regulators occur within insulated neighborhoods. These results provide new insights into the relationship between transcriptional control of cell identity genes and control of local chromosome structure.

INTRODUCTION

Embryonic stem cells depend on active transcription of genes that play prominent roles in pluripotency (ES cell identity genes) and on repression of genes encoding lineage-specifying developmental regulators (Voss et al. 2011, Orkin and Hochedlinger 2011, Young 2011). The master transcription factors (TFs) OCT4, SOX2 and NANOG (OSN) form super-enhancers at most cell identity genes, including those encoding the master TFs themselves; these super-enhancers contain exceptional levels of transcription apparatus and drive high-level expression of associated genes (Whyte et al. 2013, Hnisz et al. 2013). Maintenance of the pluripotent ESC state also requires that genes encoding lineage-specifying developmental regulators remain repressed, as expression of these genes can stimulate differentiation and thus loss of ESC identity. These repressed lineage-specifying genes are occupied by Polycomb group proteins in ESCs (Boyer et al. 2006, Lee et al. 2006, Squazzo et al. 2006, Margueron and Reinberg 2011). The ability to express or repress these key genes in a precise and sustainable fashion is thus essential to maintaining ESC identity.

Recent pioneering studies of mammalian chromosome structure have suggested that they are organized into a hierarchy of units, which include Topologically Associating Domains (TADs) and gene loops (Figure 1A)(Dixon et al. 2012, Gibcus and Dekker 2013, Nora et al. 2012, Filippova et al. 2014, Naumova et al. 2013). TADs, also known as Topological Domains, are defined by DNA-DNA interaction frequencies, and their boundaries are regions across which relatively few DNA-DNA interactions occur (Nora et al. 2012, Dixon et al.

2012). TADs average 0.8 Mb, contain approximately 7 protein-coding genes and have boundaries that are shared by the different cell types of an organism (Dixon et al. 2012, Smallwood and Ren 2013). The expression of genes within a TAD is somewhat correlated, and thus some TADs tend to have active genes and others tend to have repressed genes (Cavalli and Misteli 2013, Gibcus and Dekker 2013, Nora et al. 2012).

Gene loops and other structures within TADs are thought to reflect the activities of transcription factors (TFs), cohesin and CTCF (Phillips-Cremins et al. 2013, Zuin et al. 2014, Baranello, Kouzine, and Levens 2014, Seitan et al. 2013, Gorkin, Leung, and Ren 2014). The structures within TADs include cohesin-associated enhancer-promoter loops that are produced when enhancer-bound TFs bind cofactors such as Mediator that, in turn, bind RNA polymerase II at promoter sites (Roeder 2005, Lelli, Slattery, and Mann 2012, Spitz and Furlong 2012, Lee and Young 2013). The cohesin-loading factor NIPBL binds Mediator and loads cohesin at these enhancer-promoter loops (Kagey et al. 2010). Cohesin also becomes associated with CTCF-bound regions of the genome and some of these cohesin-associated CTCF sites facilitate gene activation while others may function as insulators (Parelho et al. 2008, Wendt et al. 2008, Dixon et al. 2012, Phillips-Cremins and Corces 2013, Seitan et al. 2013). The chromosome structures anchored by Mediator and cohesin are thought to be mostly cell-type-specific, whereas those anchored by CTCF and cohesin tend to be larger and shared by most cell types (Phillips-Cremins et al. 2013, Seitan et al. 2013). Despite this picture of cohesin-associated enhancer-promoter loops and

cohesin-associated CTCF loops, we do not yet understand the relationship between the transcriptional control of cell identity and the sub-TAD structures of chromosomes that may contribute to this control. Furthermore, there is limited evidence that that the integrity of sub-TAD structures is important for normal expression of genes located in the vicinity of these structures.

To gain insights into the cohesin-associated chromosome structures that may contribute to the control of pluripotency in ESCs, we generated a large cohesin ChIA-PET dataset and integrated this with other genome-wide data to identify local structures across the genome. The results show that super-enhancer driven cell identity genes and repressed genes encoding lineage-specifying developmental regulators occur within insulated neighborhoods formed by the looping of two CTCF interaction sites occupied by cohesin. Perturbation of these structures demonstrates that their integrity is important for normal expression of genes located in the vicinity of the neighborhoods.

RESULTS

Cohesin ChIA-PET in ESCs

The organization of mammalian chromosomes involves structural units with various sizes and properties, and cohesin, a Structural Maintenance of Chromosomes (SMC) complex, participates in DNA interactions that include enhancer-promoter loops and larger loop structures that occur within Topologically Associating Domains (TADs) (Figure 1A). ESC ChIP-seq data

indicate that ~40% of cohesin-occupied sites involve active enhancers and promoters, ~3% involve genes with Polycomb modifications, and ~50% involve CTCF sites that are not associated with enhancers, promoters or Polycomb-occupied sites (Figure 1B, S1A, S1B). We employed cohesin ChIA-PET to further investigate the relationship between control of the ESC pluripotency program and control of local chromosome structure. We selected cohesin because it is a relatively well-studied SMC complex that is loaded at enhancer-promoter loops, and can thus identify those interactions, and can also migrate to CTCF sites and thus identify those interactions as well (Kagey et al. 2010, Parelho et al. 2008, Wendt et al. 2008, Rubio et al. 2008, Schaaf et al. 2013). The ChIA-PET technique was used because it yields high-resolution (~4kb) genome-wide interaction data, which is important because most loops involved in transcriptional regulation are between 1 and 100kb (Gibcus and Dekker 2013). We hoped to extend previous findings that mapped interactions among regulatory elements across portions of the ESC genome (Phillips-Cremins et al. 2013, Seitan et al. 2013, Denholtz et al. 2013) and gain a detailed understanding of the relationship between transcriptional control of ESC identity genes and control of local chromosome structure.

To identify interactions between cohesin-occupied sites, we generated biological replicates of SMC1 ChIA-PET datasets in ESCs totaling ~400 million reads (Table S1A). The two biological replicates showed a high degree of correlation (Pearson's $r > 0.91$, Figure S1C, S1D), so we pooled the replicate data and processed it using an established protocol (Li et al. 2010), with

modifications described in Extended Experimental Procedures (Figure S1, Table S1A). The dataset contained ~19 million unique paired-end tags (PETs) that were used to identify PET peaks (Figure 1C). Interactions between PET peaks were identified and filtered for length and significance (Figure 1C, S1E, S1F, Table S1B, Extended Experimental Procedures). The analysis method produced 1,234,006 cohesin-associated DNA interactions (Figure1C, Table S1B). The vast majority (92%) of these interacting cohesin-occupied sites occurred at enhancers, promoters and CTCF binding sites, consistent with the known roles of cohesin at these regulatory elements (Figure 1D). Genomic data of any type is noisy, and our confidence in the interpretation of DNA interaction data is improved by identifying PETs that represent independent events in the sample and pass statistical significance tests. For this reason, we generated a high-confidence interaction (FDR ≤ 0.01) dataset by requiring that at least three independent PETs support the identified interaction between two PET peaks. The high-confidence dataset consisted of 23,835 interactions that were almost entirely intrachromosomal (99%), and included 2,921 enhancer-promoter interactions, 2,700 enhancer-enhancer interactions and 7,841 interactions between non-enhancer, non-promoter CTCF sites (Figure 1C, 1D, S1G, S2, Table S1B). Unless stated otherwise, the high-confidence dataset was used for further quantitative analysis.

We used the interaction datasets to create a table of enhancer-promoter assignments for ESCs (Table S2A-C). We found that the interaction data supported 83% of super-enhancer assignments to the proximal active gene and

87% of typical enhancer assignments to the proximal active gene (Table S2B, C), with approximately half of the remainder were assigned to the second most proximal gene. The interaction data most frequently assigned super-enhancers and typical enhancers to a single gene, with 76% of super-enhancers and 84% of typical enhancers showing evidence of interaction with a single gene. Prior studies have suggested there can be more frequent interactions between enhancers and genes (Kieffer-Kwon et al. 2013, Shen et al. 2012, Sanyal et al. 2012); our high-confidence data is not saturating and does not address the upper limits of these interactions (Figure S1H, Extended Experimental Procedures). The catalogue of enhancer-promoter assignments provided by these interaction data should prove useful for future studies of the roles of ESC enhancers and their associated factors in control of specific target genes.

The majority of cohesin ChIA-PET interactions did not cross the boundaries of previously defined TADs (Dixon et al. 2012, Meuleman et al. 2013, Wen et al. 2009, Filippova et al. 2014)(Figure 2, Table S3A). Figure 2A shows a representative example of a TAD, where the majority (96%) of interactions occur within the domain. As expected from previous studies, the TAD boundaries are enriched for cohesin and CTCF and thus cohesin ChIA-PET peaks (Figure 2B). Genome-wide analysis shows that 88% of all interactions are contained within TADs (Figure 2C) and are somewhat enriched near the boundaries of TADs (Figure 2D). The majority of cohesin ChIA-PET interactions did not cross lamin-associated domains (LADs), which are associated with repression at the nuclear periphery, or LOCK domains, which are large regions of chromatin marked with

histone H3K9 modifications (Table S3A) (Meuleman et al. 2013, Wen et al. 2009). These results are consistent with properties previously described for TAD, LAD and LOCK domain structures.

Super-enhancer Domain Structure

Super-enhancers drive expression of key cell identity genes and are densely occupied by the transcription apparatus and its cofactors, including cohesin (Dowen et al. 2013, Hnisz et al. 2013). Analysis of high-confidence cohesin ChIA-PET interaction data revealed a striking feature common to loci containing super-enhancers and their associated genes (Figure 3). This feature consisted of a super-enhancer and its associated gene located within a loop connected by two interacting CTCF sites co-occupied by cohesin (Figure 3A, 3B, Figure S3A-J). The vast majority of ESC super-enhancers (84%) are contained within these structures, which we call Super-enhancer Domains (SDs) (Figure 3B; Table S4A, B, Extended Experimental Procedures). In contrast, only 48% of typical enhancers were found to occur within comparable loops between two CTCF sites.

The 197 SDs average 106 kb and most frequently contain 1 or 2 genes (Table S4A, C). It was evident that there were cohesin-associated interactions between individual enhancer elements (constituents) of super-enhancers as well as interactions between super-enhancers and the promoters of their associated genes (Figure S3A-J). Indeed, the results suggest that super-enhancer constituents have cohesin-associated interactions with one another (345

interactions) even more frequently than they do with their associated genes (216 interactions).

The SDs contain high densities of pluripotency transcription factors, Mediator and cohesin, together with histone modifications associated with transcriptionally active enhancers and genes (Figure 3C). It was notable that the majority (82%) of interactions within SDs do not cross the CTCF sites at SD borders (Figure 3D) and that the majority of Mediator, Pol2 and H3K27ac signal associated with super-enhancers and their associated genes occurs inside of the CTCF sites at SD borders (Figure 3E). The cohesin ChIA-PET interaction data and the distribution of the transcription apparatus suggest that the interacting cohesin-occupied CTCF sites tend to restrict the interactions of super-enhancers to those genes within the SD.

Super-enhancer Domain Function

Because super-enhancers contain an exceptional amount of transcription apparatus and CTCF has been associated with insulator activity (Phillips and Corces 2009, Phillips-Cremins and Corces 2013, Handoko et al. 2011, Ong and Corces 2014, Essafi et al. 2011), we postulated that SD structures might be necessary for proper regulation of genes in the vicinity of these structures. To test this model, we investigated the effect of deleting SD boundary CTCF sites on expression of genes inside and immediately outside of SDs (Figure 4). For this purpose, we studied five SDs whose super-enhancer associated genes play key roles in embryonic stem cell biology (*miR-290-295, Nanog, Tdgf1, Pou5f1 (Oct4),*

and Prdm14). In all cases, we found that deletion of a CTCF site led to altered expression of nearby genes. In 4/5 cases, deletion of a CTCF site led to increased expression of genes immediately outside the SDs and in 3/5 cases, deletion of a CTCF site caused changes in expression of genes within the SDs.

The *miR-290-295* locus, which specifies miRNAs with roles in ESC biology, is located within an SD (Figure 4A). The *miR-290-295* SD contains no other annotated gene and the closest gene that resides outside this SD is *Nlrp12*, located ~20kb downstream of *miR-290-295*. CRISPR-mediated deletion of a boundary CTCF site (C1) at the *miR-290-295* locus caused a ~50% reduction in the *miR-290-295* pri-miRNA transcript and an 8-fold increase in transcript levels for *Nlrp12* (Figure 4A). The CTCF deletion had no effect on expression of two genes located further away, *AU018091* and *Myadm* (Figure 4A). These results indicate that normal expression of the *miR-290-295* pri-miRNA transcript is dependent on the CTCF boundary site and furthermore, that genes located immediately outside of this SD can be activated when the SD CTCF boundary site is disrupted.

The *Nanog* gene, which encodes a key pluripotency transcription factor, is located within an SD shown in Figure 4B. The *Nanog* SD contains no other annotated gene and the closest upstream gene that resides outside this SD is *Dppa3*, which is located ~50kb upstream of *Nanog*. CRISPR-mediated deletion of the boundary CTCF site C1 of the *Nanog* SD led to a ~40% drop in *Nanog* transcript levels (Figure 4B). In this case, there was no significant change in the

level of the *Dppa3* transcript (Figure 4B). These results indicate that normal expression of the *Nanog* transcript is dependent on the C1 CTCF site.

The *Tdgf1* gene, which encodes an epidermal growth factor essential for embryonic development, is located within an SD (Figure 4C). In this SD, it is possible that the super-enhancer regulates both the *Tdgf1* and *Lrrc2* genes and this *Tdgf1/Lrrc2* SD also contains the *Rtp3* gene. The closest gene that resides outside this SD is *Gm590*, which is located ~30kb downstream of *Tdgf1*. CRISPR-mediated deletion of a boundary CTCF site (C1) of the *Tdgf1/Lrrc2* SD had little effect on *Tdgf1* and *Rtp3* transcript levels, but had a modest effect on *Lrrc2* transcript levels and caused a nearly 10-fold increase in the levels of *Gm590* transcripts (Figure 4C).

The *Pou5f1* gene, which encodes pluripotency transcription factor OCT4, is located within an SD (Figure 4D). The *Pou5f1* SD contains no other annotated gene. We were not able to obtain a bi-allelic CRISPR-mediated deletion of a boundary CTCF site, despite multiple attempts, but did obtain a mono-allelic deletion of the boundary CTCF site C1 (Figure 4D). This mono-allelic deletion had little effect on the levels of *Pou5f1* transcripts, but increased the levels of transcripts for *H2-Q10*, the gene closest to the deleted boundary, by ~2.5-fold (Figure 4D). Transcription of the gene closest to the uninterrupted boundary of the *Pou5f1* SD, *Tcf19*, was unaffected by the C1 deletion.

The *Prdm14* gene, which encodes a pluripotency transcription factor, is located within an SD (Figure 4E). The *Prdm14* SD contains no other annotated

gene and the closest downstream gene that resides outside this SD is *Slco5a1*, which is located ~100kb downstream of *Prdm14*. The *Prdm14* SD has two neighboring cohesin-associated CTCF sites at one boundary; CRISPR-mediated deletion of a single boundary CTCF site (C1) had no effect on expression of *Prdm14* or *Slco5a1,* but deletion of both CTCF sites (C1 and C2) at that boundary caused a 3.5-fold increase in expression of *Slco5a1* (Figure 4E).

We tested whether the super-enhancers from disrupted SD structures show increased interaction frequencies with the newly activated genes outside the SD by using 3C. At two loci where loss of an SD boundary CTCF site led to significant activation of the gene outside the SD (*miR-290-295* and *Pou5f1*) we performed quantitative 3C experiments to measure the contact frequency between the super-enhancers and the genes immediately outside of SDs in wild type cells and in cells where the SD boundary CTCF site was deleted. In both cases, loss of the CTCF site led to an increase in the contact frequency between the super-enhancers and the genes immediately outside of SDs that were newly activated (Figure S4A, S4B).

We investigated whether altered SD boundaries that affect cell identity genes cause ESCs to express markers consistent with an altered cell state. Indeed, we found that ESCs lacking the *miR-290-295* boundary CTCF site C1 exhibit increased expression of the ectodermal marker *Pax6* and decreased expression of the endodermal lineage markers *Gata6* and *Sox17*, suggesting that loss of the SD structure is sufficient to affect cell identity (Figure S4C). Previous studies have shown that *miR-290-295* null ESCs show an increased propensity

to differentiate into ectodermal lineages at the expense of endoderm (Kaspi et al. 2013).

In summary, the loss of CTCF sites at the boundaries of SDs can cause a change in the level of transcripts for super-enhancer associated genes within the SD and frequently leads to activation of genes near these CTCF sites. These results indicate that the integrity of SDs is important for normal expression of genes located in the vicinity of the SD, which can include genes that are key to control of cell identity.

Polycomb Domains

Maintenance of the pluripotent ESC state requires that genes encoding lineage-specifying developmental regulators are repressed, and these repressed lineage-specifying genes are occupied by nucleosomal histones that carry the Polycomb-associated mark H3K27me3 (Young 2011, Margueron and Reinberg 2011). The mechanisms responsible for maintaining the H3K27me3 mark across short spans of regulatory regions and promoters of repressed genes are not well understood, although CTCF sites have been implicated (Schwartz et al. 2012, Van Bortle et al. 2012, Cuddapah et al. 2009). Analysis of the H3K27me3-marked genes revealed that they, like the super-enhancer-associated genes, are typically located within a loop between two interacting CTCF sites co-occupied by cohesin (Figure 5A, 5B, Figure S5A-J, Table S5A). These Polycomb Domain (PD) structures share many features with the Super-enhancer Domains. The majority (70%) (380/546) of Polycomb-associated genes occur in PD structures.

PDs average 112 kb and generally contain 1 or 2 genes (Table S5B). The PDs contain exceptionally high densities of the Polycomb proteins EZH2, SUZ12 and the associated histone modification H3K27me3 (Figure 5C). The majority (78%) of cohesin ChIA-PET interactions originating in PDs occur within the PD boundaries (Figure 5D). Furthermore, the Polycomb mark H3K27me3 tends to be retained within the PD (Figure 5E).

We postulated that the CTCF boundaries that form PD structures might be important for repression of the Polycomb-marked genes within the PD, and investigated the effect of deleting boundary CTCF sites on a PD containing *Tcfap2e* to test this idea (Figure 5F). CRISPR-mediated deletion of one of the boundary CTCF sites (C1) of the *Tcfap2e* PD caused a 1.7 fold increase in transcript levels for *Tcfap2e* (P-value < 0.05) and no significant change in transcript levels for nearby genes within or outside of the PD. CRISPR-mediated deletion of the other boundary CTCF site (C2) caused a 4-fold increase in the expression of *Tcfap2e* (P-value < 0.001) and little effect on adjacent genes. These results suggest that the integrity of the CTCF boundaries of PDs is important for full repression of H3K27me3-occupied genes.

Insulated Neighborhoods in Multiple Cell Types

A previous study suggested that DNA loops mediated by cohesin and CTCF tend to be larger and more shared among multiple cell types than DNA loops associated with cohesin and Mediator, which represent enhancer-promoter interactions that may be cell-type specific (Phillips-Cremins et al. 2013). This led

us to postulate that 1) the interacting CTCF structures of SDs and PDs may be common to multiple cell types, and 2) the acquisition of super-enhancers and Polycomb binding within these common domain structures will vary based on the gene expression program of the cell type (Figure 6A).

To test this model, we compared the SDs identified in ESCs to comparable regions in neural precursor cells (NPCs) where 5C interaction data was available for specific loci (Phillips-Cremins et al. 2013). We found, for example, that the *Nanog* locus SD observed in ESCs with ChIA-PET data was also detected by 5C data in NPCs (Figure 6B). In NPCs, the *Nanog* gene is not expressed and no super-enhancers are formed at this locus (Figure 6B). Similarly, there is evidence for a common structure involving CTCF sites bounding the *Olig1/Olig2* locus in both ESCs and NPCs (Figure 6B). In this domain, the *Olig1/Olig2* genes are not active and no super-enhancers are formed in ESCs, whereas there are three super-enhancers in NPCs, where these genes are highly expressed (Figure 6B, S6A). For regions where 5C interaction data in NPCs and ChIA-PET interaction data in ESCs could be compared, a total of 11 out of 32 interactions between CTCF sites identified in NPCs were supported by interaction data in ESCs (Table S3B), which is impressive given the sparsity of interaction data. This supports the view that the interacting CTCF structures of ESC SDs may be common to multiple cell types.

If the CTCF boundaries of ESC SDs and PDs are common to many cell types, we would expect that the binding of CTCF to the SD and PD boundary sites observed in ESCs will be conserved across multiple cell types. To test this

notion, we examined CTCF ChIP-seq peaks from 18 mouse cell types and determined how frequently CTCF binding occurred across these cell types (Figure 6C). When all ESC CTCF ChIP-seq peaks were included in the analysis, we found that there was fairly even distribution of the data into bins representing one or more cell types (Figure 6C). In contrast, CTCF peaks co-bound by cohesin, which included those at SD and PD borders were observed more frequently in bins representing a larger fraction of the cell types (Figure 6C; Figure S6B). These results indicate that the CTCF boundary sites of ESC SDs and PDs are frequently occupied by CTCF in multiple cell types, and together with the analysis of interaction data for NPCs described above, support the idea that CTCF-CTCF interaction structures may often be shared by ESCs and more differentiated cell types.

DISCUSSION

Understanding how the ESC pluripotency gene expression program is regulated of considerable interest because it provides the foundation for understanding gene control in all cells. There is much evidence that cohesin and CTCF have roles in connecting gene regulation and chromosome structure in ESCs (Dixon et al. 2012, Merkenschlager and Odom 2013, Phillips-Cremins and Corces 2013, Phillips-Cremins et al. 2013, Sanyal et al. 2012, Gorkin, Leung, and Ren 2014, Sofueva et al. 2013, Cavalli and Misteli 2013, Gibcus and Dekker 2013) but limited knowledge of the these structures across the genome and

scant functional evidence that specific structures actually contribute to the control of important ESC genes. We describe here organizing principles that explain how a key set of cohesin-associated chromosome structures contribute to the ESC gene expression program.

To gain insights into the relationship between transcriptional control of cell identity and control of chromosome structure, we carried out cohesin ChIA-PET and focused the analysis on loci containing super-enhancers, which drive expression of key cell identity genes. We found that the majority of super-enhancers and their associated genes occur within large loops that are connected through interacting CTCF sites co-occupied by cohesin. These super-enhancer domains, or SDs, typically contain one super-enhancer that loops to one gene within the SD. The SDs appear to restrict super-enhancer activity to genes within the SD, because the cohesin ChIA-PET interactions occur primarily within the SD and loss of a CTCF boundary tends to cause inappropriate activation of nearby genes located outside that boundary. The proper association of super-enhancers and their target genes in such "insulated neighborhoods" is of considerable importance since the mis-targeting of a single super-enhancer is sufficient to cause leukemia (Groschel et al. 2014).

The cohesin ChIA-PET data and perturbation of CTCF sites suggest that genes that encode repressed, lineage-specifying, developmental regulators also occur within insulated neighborhoods in ESCs. Maintenance of the pluripotent ESC state requires that genes encoding lineage-specifying developmental regulators are repressed, and these repressed lineage-specifying genes are

occupied by nucleosomal histones that carry the Polycomb mark H3K27me3 (Lee et al. 2006, Boyer et al. 2006, Schwartz et al. 2006, Tolhuis et al. 2006, Squazzo et al. 2006, Negre et al. 2006, Bracken et al. 2006). The majority of these genes were found to be located within a cohesion-associated CTCF-CTCF loop, which we call a Polycomb Domain, or PD. The perturbation of CTCF PD boundary sites caused de-repression of the Polycomb-bound gene within the PD, suggesting that these boundaries are important for maintenance of gene repression within the PD.

CTCF has previously been shown to be associated with boundary formation, insulator activity and transcriptional regulation (Handoko et al. 2011, Denholtz et al. 2013, Sexton et al. 2012, Schwartz et al. 2012, Phillips and Corces 2009, Felsenfeld et al. 2004, Valenzuela and Kamakaka 2006, Bell, West, and Felsenfeld 1999, Kim et al. 2007, Soshnikova et al. 2010). Previous reports have also demonstrated that cohesin and CTCF are associated with large loop substructures within TADs, whereas cohesin and Mediator are associated with smaller loop structures that sometimes form within the CTCF-bounded loops (Sofueva et al. 2013, Phillips-Cremins et al. 2013, de Wit et al. 2013). CTCF-bound domains have been proposed to confine the activity of enhancers to specific target genes, thus yielding proper tissue-specific expression of genes (Hawkins et al. 2011, Demare et al. 2013, Handoko et al. 2011). Our genome-wide study extends these observations by connecting such structures with the transcriptional control of specific super-enhancer-driven and Polycomb-repressed cell identity genes, and by showing that these structures can contribute to the

control of genes inside and outside of the insulated neighborhoods that contain key pluripotency genes.

The organization of key cell identity genes into insulated neighborhoods may be a property common to all mammalian cell types. Indeed, several recent studies have identified CTCF bounded regions whose function is consistent with ESC SDs (Wang et al. 2014, Guo et al. 2011). For example, in T cell acute lymphocytic leukemia, Notch1 activation leads to increased expression of a super-enhancer--driven gene found between two CTCF sites that are structurally connected, but does not affect genes located outside of the two CTCF sites (Wang et al. 2014). Future studies addressing the mechanisms that regulate loop formation should provide additional insights into the relationships between transcriptional control of cell identity genes and control of local chromosome structure.

EXPERIMENTAL PROCEDURES

Cell Culture

V6.5 murine ESCs were grown on irradiated murine embryonic fibroblasts (MEFs) under standard ESC conditions as described previously (Whyte et al. 2012).

Genome Editing

The CRISPR/Cas9 system was used to create ESC lines with CTCF site deletions. Target-specific oligonucleotides were cloned into a plasmid carrying a codon-optimized version of Cas9 (pX330, Addgene: 42230). The genomic sequences complementary to guide RNAs in the genome editing experiments are listed in the Extended Experimental Procedures. Cells were transfected with two plasmids expressing Cas9 and sgRNA targeting regions around 200 basepairs up- and down- stream of the CTCF binding site, respectively. A plasmid expressing PGK-puroR was also co-transfected, using X-fect reagent (Clontech) according to the manufacturer's instructions. One day after transfection, cells were re-plated on DR4 MEF feeder layers. One day after re-plating, puromycin (2ug/ml) was added for three days. Subsequently, puromycin was withdrawn for three to four days. Individual colonies were picked and genotyped by PCR.

ChIA-PET

SMC1 ChIA-PET was performed as previously described (Chepelev et al. 2012, Fullwood et al. 2009, Li et al. 2012, Goh et al. 2012). Briefly, murine ESCs (up to $1x10^8$ cells) were treated with 1% formaldehyde at room temperature for 10 min and then neutralized using 0.2M glycine. The crosslinked chromatin was fragmented by sonication to size lengths of 300-700 bp. The anti-SMC1 antibody (Bethyl, A300-055A) was used to enrich SMC1-bound chromatin fragments. A portion of ChIP DNA was eluted from antibody-coated beads for concentration quantification and for enrichment analysis using quantitative PCR. For ChIA-PET

library construction ChIP DNA fragments were end-repaired using T4 DNA polymerase (NEB) and ligated to either linker A or linker B. After linker ligation, the two samples were combined for proximity ligation in diluted conditions. Following proximity ligation, the Paired-End Tag (PET) constructs were extracted from the ligation products and the PET templates were subjected to 50x50 paired-end sequencing using Illumina HiSeq 2000.

Data analysis

ChIA-PET data analysis was performed as previously described (Li et al. 2010), with modifications described in the Extended Experimental Procedures. The high confidence interactions for the two biological replicate SMC1 ChIA-PET experiments and for the merged dataset are listed in Tables S1C, S1D and S1E, respectively. All datasets used in this study are listed in Table S6.

ACCESSION NUMBERS

Raw and processed sequencing data were deposited in GEO under accession number GSE57913 (www.ncbi.nlm.nih.gov/geo/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, 6 Figures and 6 Tables and can be found with this article online.

AUTHOR CONTRIBUTIONS

J.M.D and G.R. performed ChIA-PET. Z.F.P performed ChIA-PET data analysis with help from L.N.Z. Genome-wide computational analyses were

performed by Z.F.P., B.J.A. and L.N.Z. D.H and A.S.W. designed and performed genome editing experiments. D.H., A.S.W. and J.S. performed gene expression analyses. G.R. performed 3C experiments. T.I.L and K.Z. contributed to the conceptual development of the study. J.M.D., Z.P.F, D.H and R.A.Y wrote the paper. All authors edited the manuscript.

ACKNOWLEDGEMENTS

We thank Warren Whyte for generating the H3K27me3 ChIP-Seq dataset, Chikdu Shivalila for help with genome editing experiments, Alla Sigova for help with experiment design, Rudolf Jaenisch for sharing CRISPR reagents, and the Whitehead Institute Genome Technology Core and NHLBI DNA Sequencing Core for Illumina sequencing. We also thank members of the Young lab for helpful discussions. This work was supported by the National Institutes of Health grant HG002668 (R.A.Y.), Division of Intramural Research, NHLBI (K.Z.), a Ruth L. Kirschstein National Research Service Award (CA168263-01A1) (J.M.D.), by an Erwin Schrödinger Fellowship (J3490) from the Austrian Science Fund (FWF) (D.H.), and a Rubicon Fellowship for the Life Sciences, Netherlands Organization for Scientific Research (NWO)(J.S.). R.A.Y. is a founder of Syros Pharmaceuticals.

FIGURES

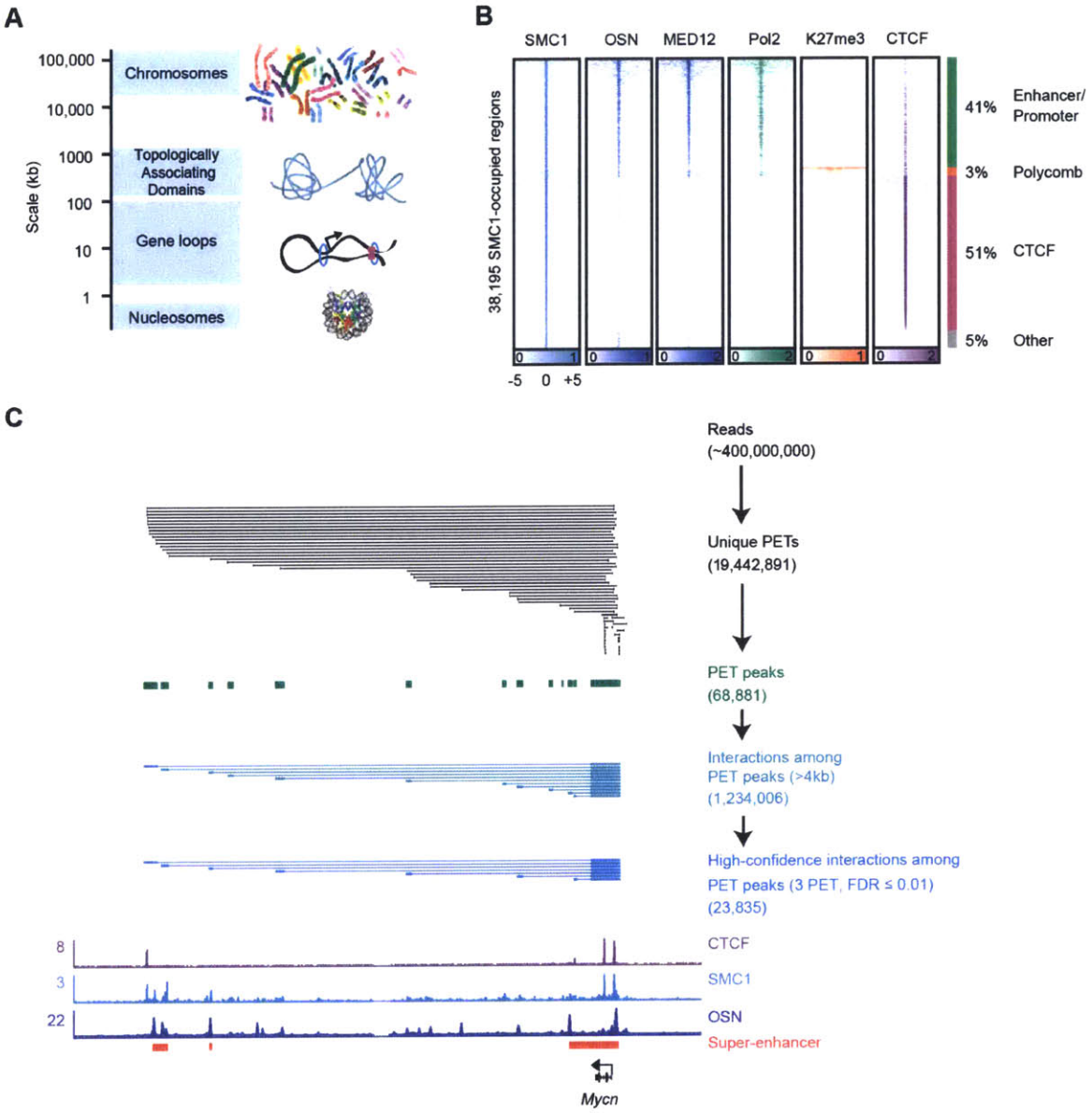

A
100,000
10,000
1000
100
10
1
Scale (kb)
Chromosomes
Topologically Associating Domains
Gene loops
Nucleosomes
B
SMC1
OSN
MED12
Pol2
K27me3
CTCF
38,195 SMC1-occupied regions
-5 0 +5
41% Enhancer/ Promoter
3% Polycomb
51% CTCF
5% Other
C
Reads (~400,000,000)
Unique PETs (19,442,891)
PET peaks (68,881)
Interactions among PET peaks (>4kb) (1,234,006)
High-confidence interactions among PET peaks (3 PET, FDR ≤ 0.01) (23,835)
8
3
22
CTCF
SMC1
OSN
Super-enhancer
Mycn
D
Interactions
CTCF-CTCF (262,149)
CTCF site (28,845)
Enhancer-CTCF (305,762)
Promoter-CTCF (229,708)
Enhancer (18,226)
Promoter (14,232)
Enhancer-Enhancer (104,394)
Promoter-Promoter (56,101)
Enhancer-Promoter (148,080)
High-confidence interactions (FDR ≤ 0.01, 3 PETs)
CTCF-CTCF (7,841)
CTCF site (13,430)
Enhancer-CTCF (5,358)
Promoter-CTCF (2,921)
Enhancer (5,942)
Promoter (3,404)
Enhancer-Enhancer (2,700)
Promoter-Promoter (818)
Enhancer-Promoter (2,921)

Figure 1. DNA interactions involving cohesin.

A) Units of chromosome organization. Chromosomes consist of multiple Topologically Associating Domains (TADs). TADs (image adapted from (Dixon et al. 2012)) contain multiple genes with DNA loops involving interactions between enhancers, promoters and other regulatory elements, which are mediated by cohesin (blue ring) and CTCF (purple balls). Nucleosomes represent the smallest unit of chromosome organization.

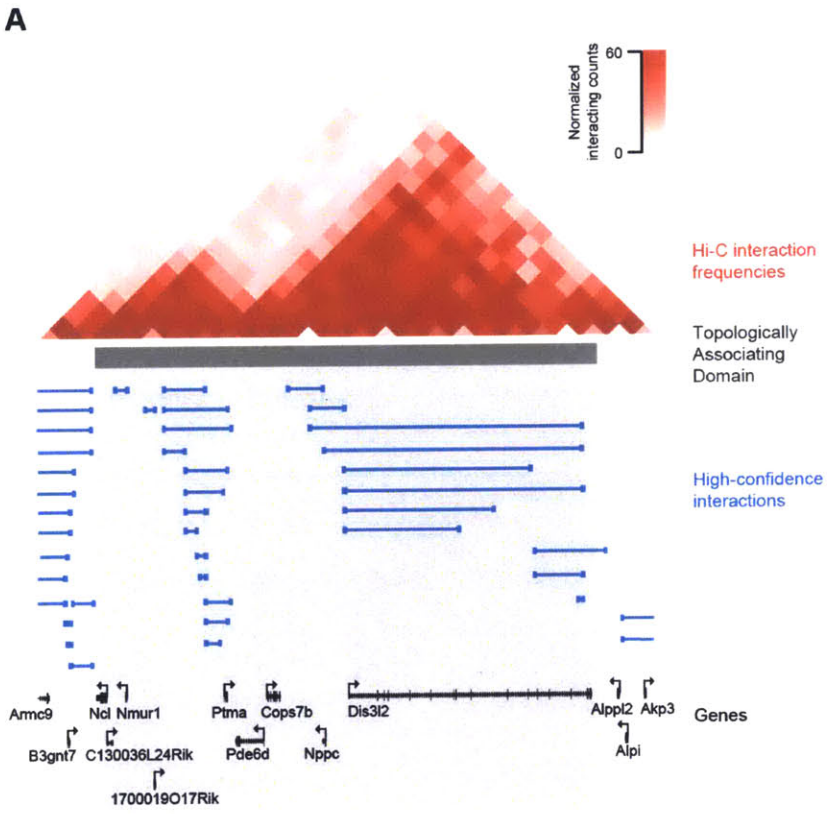
B) Heatmap representation of ESC ChIP-seq data for SMC1, a merged dataset for the transcription factors OCT4, SOX2 and NANOG (OSN), MED12, RNA polymerase II (Pol2), H3K27me3, and CTCF at SMC1-occupied regions. Read density is displayed within a 10kb window and color scale intensities are shown in rpm/bp. Cohesin occupies three classes of sites: enhancer-promoter sites, Polycomb-occupied sites, and CTCF-occupied sites.

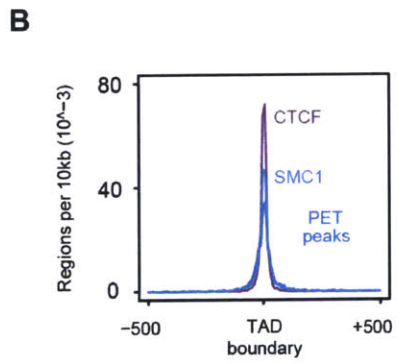
C) ESC cohesin (SMC1) ChIA-PET data analysis at the *Mycn* locus. The algorithm used to identify paired-end tags (PETs) is described in detail in Extended Experimental Procedures. PETs and interactions involving enhancers and promoters within the window are displayed at each step in the analysis pipeline: unique PETs, PET peaks, interactions between PET peaks, and high-confidence interactions supported by at least 3 independent PETs and with a FDR of 0.01.

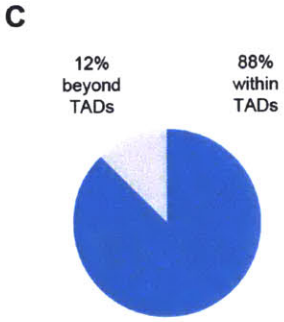
D) Summary of the major classes of interactions and high-confidence interactions

identified in the cohesin ChIA-PET data. Enhancers, promoters, and CTCF sites where interactions occur are displayed as blue circles, and the size of the circle is proportional to the number regions. The interactions between two sites are displayed as grey lines, and the thickness of the grey line is proportional to the number of interactions. The diagram on the left was generated using the interactions, and the diagram on the right was generated using the high confidence interactions.

See also Figure S1, S2, Table S1, S2.


A
Normalized interacting counts
60
0
Hi-C interaction frequencies
Topologically Associating Domain
High-confidence interactions
Armc9
Ncl
Nmur1
Ptma
Cops7b
Dis3l2
Alppl2
Akp3
Genes
B3gnt7
C130036L24Rik
Pde6d
Nppc
Alpi
1700019O17Rik


B
Regions per 10kb (10^-3)
80
40
0
CTCF
SMC1
PET peaks
−500
TAD boundary
+500


C
12% beyond TADs
88% within TADs

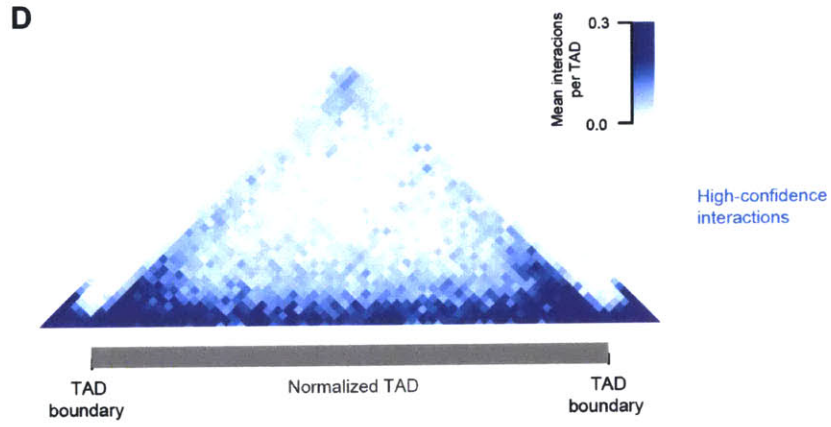

D
Mean interacions per TAD
0.3
0.0
High-confidence interactions
TAD boundary
Normalized TAD
TAD boundary

Figure 2. DNA interactions frequently occur within Topologically Associating Domains.

A) An example Topologically Associating Domain (TAD) shown with normalized Hi-C interaction frequencies displayed as a two-dimensional heat map (Dixon et al. 2012) and the TAD is indicated as a grey bar. High-confidence SMC1 ChIA-PET interactions are depicted as blue lines.

B) Enrichment of CTCF, cohesin (SMC1), and PET peaks at TAD boundary regions. The metagene representation shows the number of regions per 10 kb window centered on the TAD boundary and +/- 500kb is displayed.

C) Pie chart of high-confidence interactions that either fall within TADs (88%) or cross TAD boundaries (12%).

D) High-confidence interactions are displayed as a two-dimensional heat map across a normalized TAD length for the ~2,200 TADs (Dixon et al. 2012). The display is centered on the normalized TAD and extends beyond each boundary to 10% of the size of the domain.

See also Table S3A.

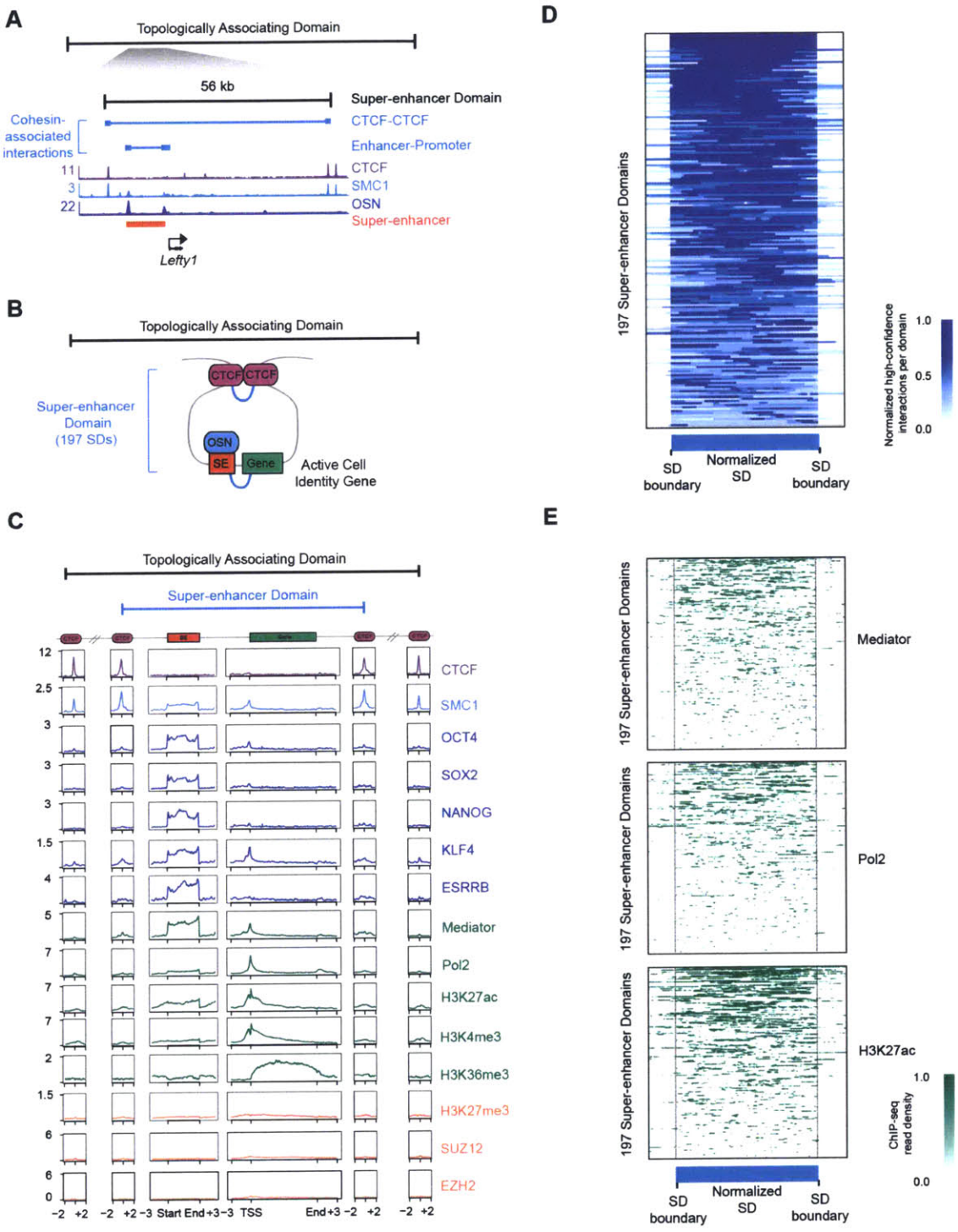

A
Topologically Associating Domain
56 kb
Super-enhancer Domain
Cohesin-associated interactions
CTCF-CTCF
Enhancer-Promoter
11
3
22
CTCF
SMC1
OSN
Super-enhancer
Lefty1
B
Topologically Associating Domain
Super-enhancer Domain (197 SDs)
CTCF
CTCF
OSN
SE
Gene
Active Cell Identity Gene
C
Topologically Associating Domain
Super-enhancer Domain
12
2.5
3
3
3
1.5
4
5
7
7
7
2
1.5
6
6
0
CTCF
SMC1
OCT4
SOX2
NANOG
KLF4
ESRRB
Mediator
Pol2
H3K27ac
H3K4me3
H3K36me3
H3K27me3
SUZ12
EZH2
−2 +2 −2 +2 −3 Start End +3 −3 TSS End +3 −2 +2 −2 +2
D
197 Super-enhancer Domains
SD boundary
Normalized SD
SD boundary
Normalized high-confidence interactions per domain
1.0
0.5
0.0
E
197 Super-enhancer Domains
Mediator
197 Super-enhancer Domains
Pol2
197 Super-enhancer Domains
H3K27ac
SD boundary
Normalized SD
SD boundary
ChIP-seq read density
1.0
0.0

Figure 3. Super-enhancer Domain Structure.

A) An example super-enhancer domain (SD) within a TAD. High-confidence SMC1 ChIA-PET interactions are depicted as blue lines. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and the master transcription factors OCT4, SOX2, and NANOG (OSN) are shown at the *Lefty1* locus in ESCs. The super-enhancer is indicated by a red bar.

B) Model of SD structure. The 197 SDs have interactions (blue) between cohesin-occupied CTCF sites that may serve as outer boundaries of the domain structure. SDs also contain interactions between super-enhancers and the promoters of their associated genes.

C) Metagene analysis showing the occupancy of various factors at the key elements of TADs and SDs, including CTCF sites, super-enhancers and super-enhancer associated genes. ChIP-seq profiles are shown in reads per million per base pair. Boundary site metagenes are centered on the CTCF peak, and +/-2kb is displayed. Super-enhancer metadata is centered on the 195 super-enhancers in SDs and +/-3 kb is displayed. The data for associated genes are centered on the 219 super-enhancer -associated genes in SDs and +/-3kb is displayed.

D) Heat map showing that cohesin ChIA-PET high-confidence interactions occur predominantly within the SDs. The density of high-confidence interactions is

shown across a normalized SD length for the 197 SDs.

E) Heat màp showing that transcriptional proteins are contained within boundary sites of SDs. The occupancy of Mediator (MED12), H3K27ac and RNA polymerase II (Pol2) at super-enhancers and associated genes is shown across a normalized SD length for the 197 SDs.

See also Figure S3, Table S4.

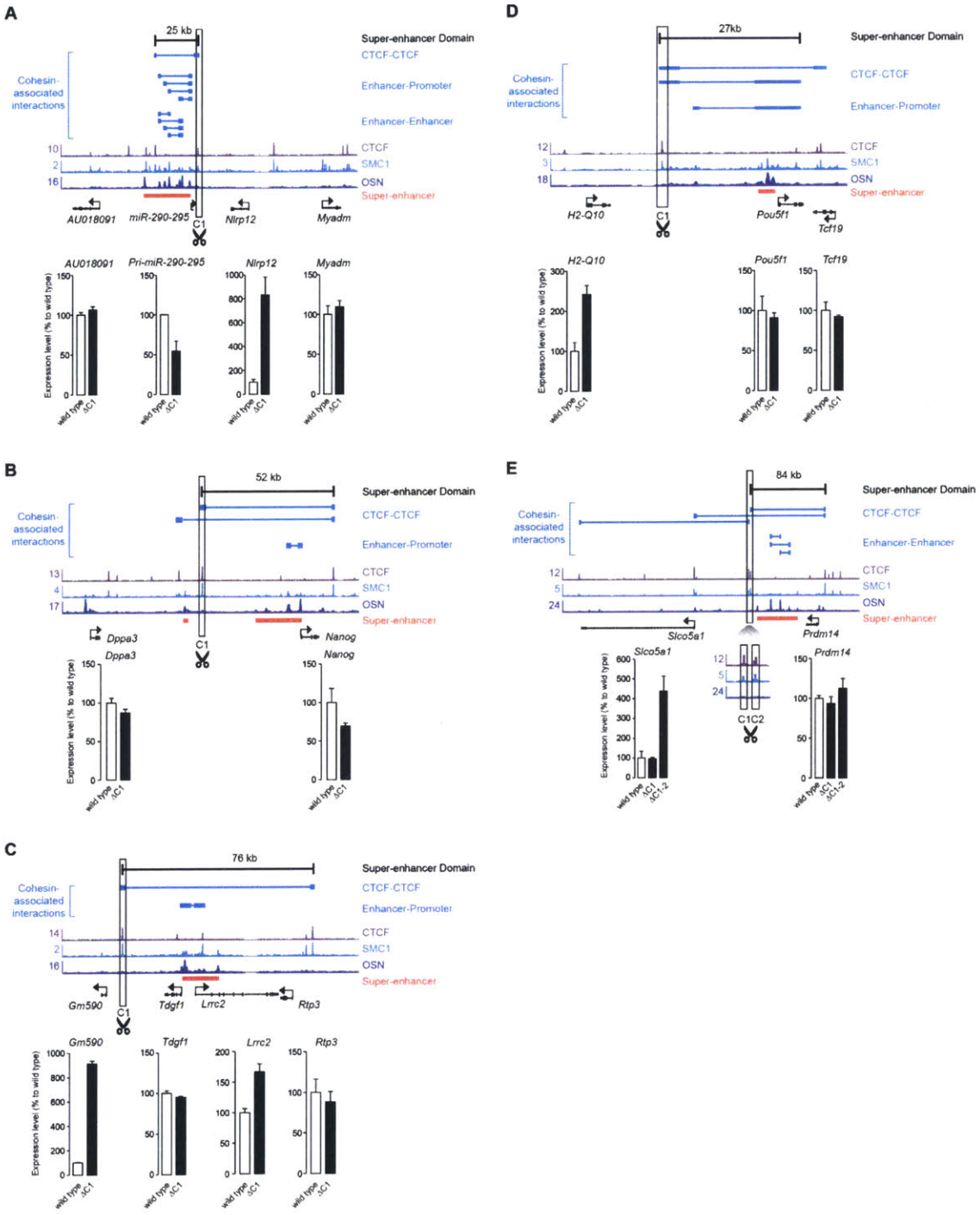


Figure 4. Super-enhancer Domains are functionally linked to gene expression.

CRISPR-mediated genome editing of CTCF sites at five loci. The top of each panel shows high-confidence interactions depicted as blue lines, and ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and OCT4, SOX2, and NANOG (OSN) in ESCs at the respective loci. The super-enhancer is indicated as a red bar. The bottom of each panel shows gene expression level of the indicated genes in wild type and CTCF site-deleted cells measured by qRT-PCR. Transcript levels were normalized to *GAPDH*. Gene expression was assayed in triplicate in at least two biological replicate samples, and is displayed as mean+SD. All P-values were determined using the Student´s t-test.

A) CRISPR-mediated genome editing of a CTCF site at the *miR-290-295* locus. (P-value < 0.001, *Pri-miR-290-295* and *Nlrp12* in wild-type vs. CTCF site-deleted).

B) CRISPR-mediated genome editing of a CTCF site at the *Nanog* locus. (P-value < 0.05, *Nanog* in wild-type vs. CTCF site-deleted).

C) CRISPR-mediated genome editing of a CTCF site at the *Tdgf1* locus. (P-value < 0.001, *Gm590;* P-value < 0.01, *Lrrc2)* in wild-type vs. CTCF site-deleted).

D) CRISPR-mediated genome editing of a CTCF site at the *Pou5f1* locus. (P-value < 0.012, *H2Q-10* in wild-type vs. CTCF site-deleted).

E) CRISPR-mediated genome editing of CTCF sites at the *Prdm14* locus. (P-value < 0.001, *Slco5a1* in wild-type vs. CTCF site-deleted).

The CTCF-deletion lines at the *Pou5f1* and *Prdm14* (C1-2) loci are heterozygous, while the CTCF-deletion lines at the *Nanog*, *Tdgf1* and *miR-290-295* loci are homozygous for the mutation.

See also Figure S4.

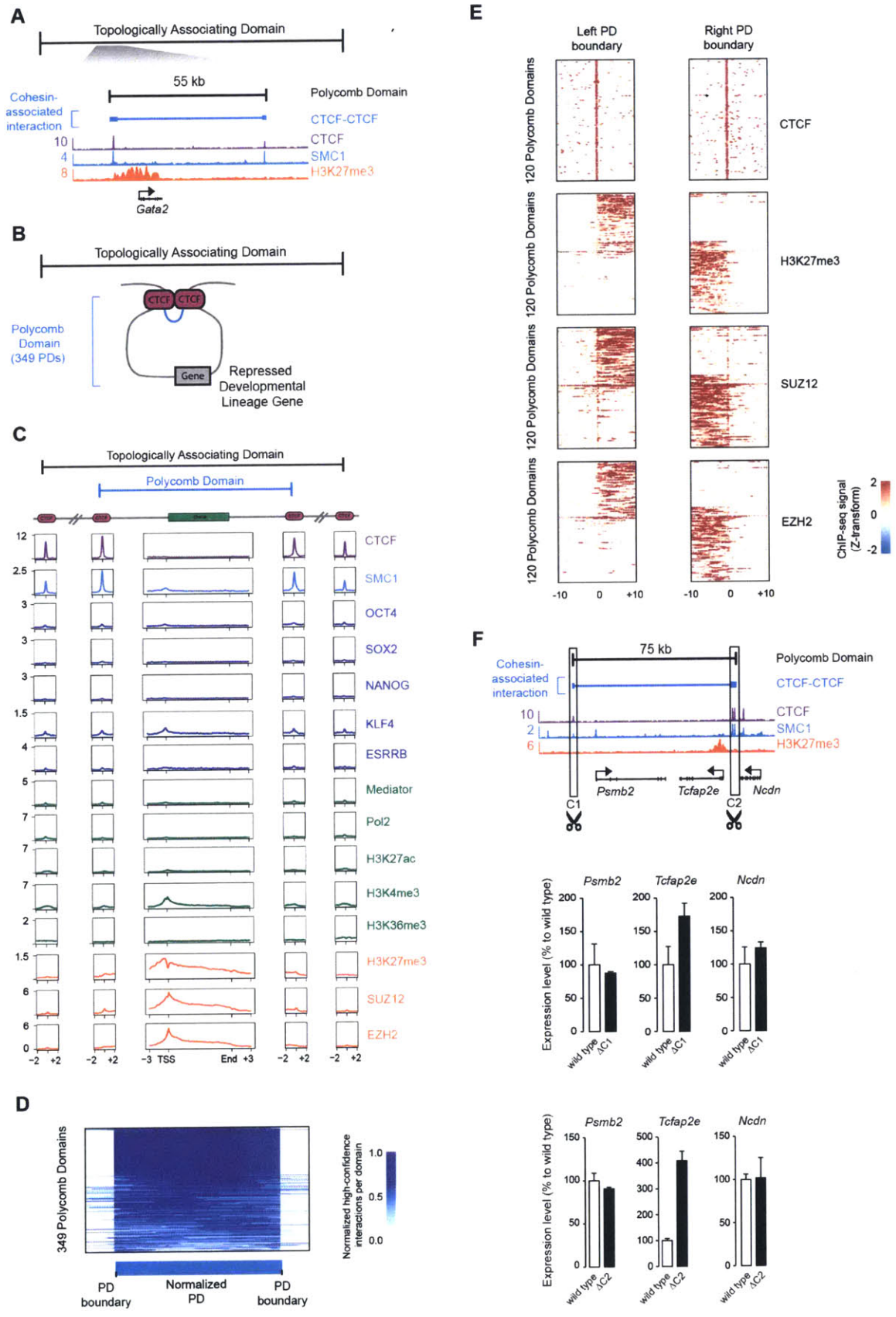

A
Topologically Associating Domain
55 kb
Polycomb Domain
Cohesin-associated interaction
CTCF-CTCF
CTCF
SMC1
H3K27me3
Gata2
B
Topologically Associating Domain
Polycomb Domain (349 PDs)
CTCF
Gene
Repressed Developmental Lineage Gene
C
Topologically Associating Domain
Polycomb Domain
CTCF
SMC1
OCT4
SOX2
NANOG
KLF4
ESRRB
Mediator
Pol2
H3K27ac
H3K4me3
H3K36me3
H3K27me3
SUZ12
EZH2
TSS
End
D
349 Polycomb Domains
Normalized high-confidence interactions per domain
PD boundary
Normalized PD
PD boundary
E
Left PD boundary
Right PD boundary
120 Polycomb Domains
CTCF
H3K27me3
SUZ12
EZH2
ChIP-seq signal (Z-transform)
F
75 kb
Polycomb Domain
Cohesin-associated interaction
CTCF-CTCF
CTCF
SMC1
H3K27me3
Psmb2
Tcfap2e
Ncdn
C1
C2
Expression level (% to wild type)
wild type
ΔC1
ΔC2

Figure 5. Polycomb Domain Structure.

A) An example Polycomb Domain (PD) within a TAD. A high-confidence interaction is depicted as the blue line. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and H3K27me3 at the *Gata2* locus in ESCs.

B) Model of PD structure. The 349 PDs have interactions (blue) between CTCF sites that serve as putative boundaries of the domain structure.

C) Metagene analysis reveals the occupancy of various factors at the key elements of TADs and PDs: CTCF sites and target genes. ChIP-seq profiles are shown in reads per million per base pair. Boundary site metagenes are centered on the CTCF peak and +/-2 kb is displayed. The metagenes depicting genes are centered on the 380 Polycomb target genes in PDs and +/-3 kb is displayed.

D) Heat map showing that high-confidence interactions are largely constrained within PDs. The density of high-confidence interactions is shown across a normalized PD length for the 349 PDs.

E) Heat map showing that Polycomb proteins are contained within boundary sites of PDs. The occupancy of CTCF, H3K27me3, SUZ12 and EZH2 is indicated within a 10 kb window centered on the left and right CTCF-occupied boundary regions is shown for the 120 PDs with this transition pattern.

F) CRISPR-mediated genome editing of a CTCF site at the *Tcfa2e* locus. *Top*, high-confidence interactions are depicted by blue lines and ChIP-Seq binding

profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and H3K27me3 are shown in ESCs. *Bottom,* Expression level of the indicated genes in wild type and CTCF site-deleted cells measured by qRT-PCR. Transcript levels were normalized to *GAPDH*. Gene expression was assayed in triplicate in at least two biological replicate samples and is displayed as mean+SD (P-value < 0.05, *Tcfap2e* in C1 deletion cells*;* P-value < 0.001, *Tcfap2e* in C2 deletion cells*)* in wild-type vs. CTCF site-deleted). P-values were determined using the Student´s t-test.

See also Figure S5, Table S5.

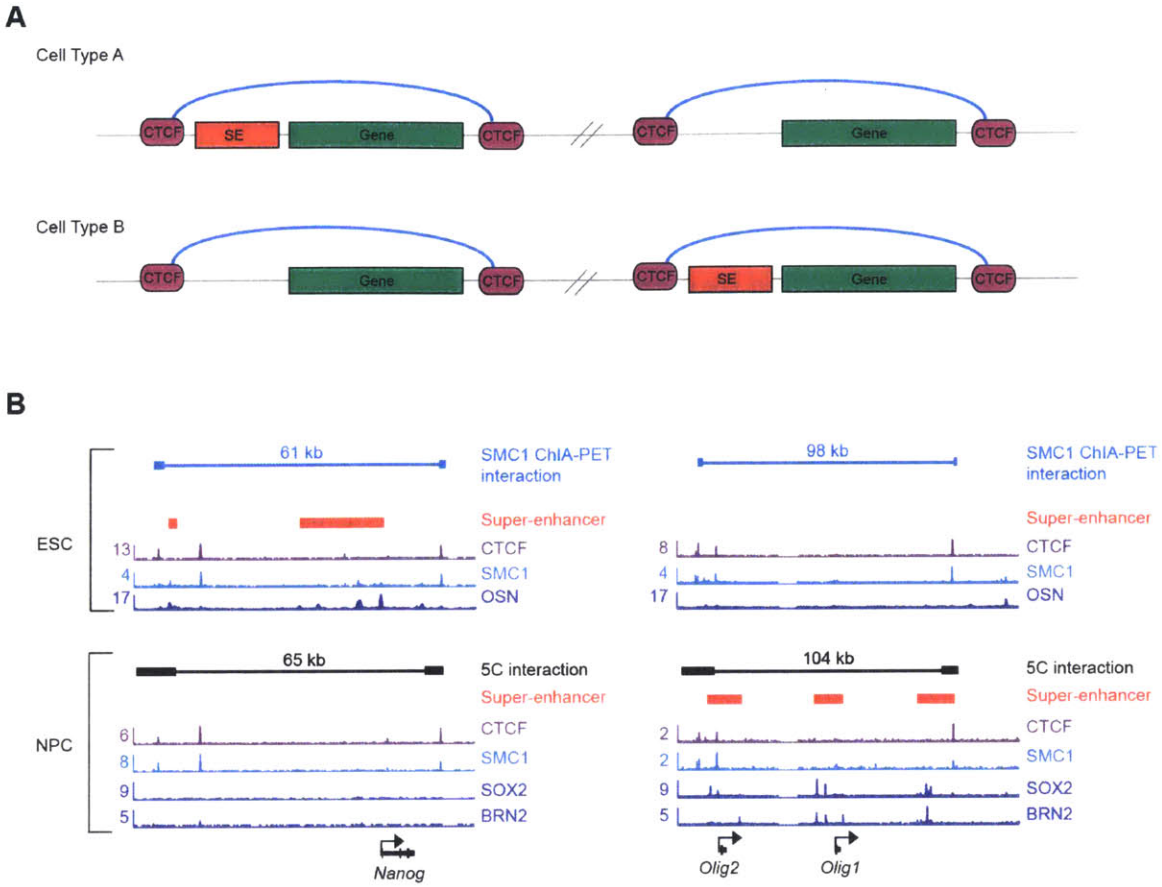

A
Cell Type A
CTCF
SE
Gene
CTCF
CTCF
Gene
CTCF
Cell Type B
CTCF
Gene
CTCF
CTCF
SE
Gene
CTCF
B
ESC
61 kb
SMC1 ChIA-PET interaction
Super-enhancer
13
CTCF
4
SMC1
17
OSN
98 kb
SMC1 ChIA-PET interaction
Super-enhancer
8
CTCF
4
SMC1
17
OSN
NPC
65 kb
5C interaction
Super-enhancer
6
CTCF
8
SMC1
9
SOX2
5
BRN2
Nanog
104 kb
5C interaction
Super-enhancer
2
CTCF
2
SMC1
9
SOX2
5
BRN2
Olig2
Olig1
C
% of the CTCF ChIP-Seq peaks identified in ESCs
all CTCF peaks observed in ESC
CTCF peaks at SD borders
CTCF peaks at PD borders
CTCF peaks at PET peaks
20
10
0
1
(specific)
18
(constitutive)
Number of cell types in which the CTCF ChIP-Seq peak is observed

Figure 6. Insulated Neighborhoods are preserved in multiple cell types.

A) Model depicting constitutive domain organization, mediated by interaction of two CTCF sites co-occupied by cohesin, in two cell types.

B) An example SD in ESCs and a domain in NPCs. High-confidence interactions from the SMC1 ChIA-PET dataset are depicted by blue lines and 5C interactions from (Phillips-Cremins et al. 2013) are depicted by black lines. Super-enhancers are indicated by red bars. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and OCT4, SOX2, and NANOG (OSN), SOX2 and BRN2 are shown at the *Nanog* locus and the *Olig1/Olig2* locus in ESCs and NPCs.

C) Occupancy of CTCF peaks across 18 cell types. The CTCF peaks used for the analysis are the CTCF peaks found in ESCs. The percentage of these peaks that are observed in the indicated number of cell types is shown for four groups of CTCF sites: all CTCF peaks identified in ESCs, CTCF peaks at SD boundaries in ESCs, CTCF peaks at PD boundaries in ESCs, and CTCF peaks at PET peaks (identified by SMC1 ChIA-PET in ESCs).

See also Figure S6, Table S3B.

REFERENCES

Baranello, L., F. Kouzine, and D. Levens. 2014. "CTCF and cohesin cooperate to organize the 3D structure of the mammalian genome." *Proc Natl Acad Sci U S A* 111 (3):889-90. doi: 10.1073/pnas.1321957111.

Bell, A. C., A. G. West, and G. Felsenfeld. 1999. "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." *Cell* 98 (3):387-96.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch. 2006. "Polycomb complexes repress developmental regulators in murine embryonic stem cells." *Nature* 441 (7091):349-53. doi: nature04733 [pii] 10.1038/nature04733.

Bracken, A. P., N. Dietrich, D. Pasini, K. H. Hansen, and K. Helin. 2006. "Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions." *Genes Dev* 20 (9):1123-36. doi: 10.1101/gad.381706.

Cavalli, G., and T. Misteli. 2013. "Functional implications of genome topology." *Nat Struct Mol Biol* 20 (3):290-9. doi: nsmb.2474 [pii] 10.1038/nsmb.2474.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang, and K. Zhao. 2012. "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization." *Cell Res* 22 (3):490-503. doi: 10.1038/cr.2012.15.

Cuddapah, S., R. Jothi, D. E. Schones, T. Y. Roh, K. Cui, and K. Zhao. 2009. "Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains." *Genome Res* 19 (1):24-32. doi: 10.1101/gr.082800.108.

de Wit, E., B. A. Bouwman, Y. Zhu, P. Klous, E. Splinter, M. J. Verstegen, P. H. Krijger, N. Festuccia, E. P. Nora, M. Welling, E. Heard, N. Geijsen, R. A. Poot, I. Chambers, and W. de Laat. 2013. "The pluripotent genome in three dimensions is shaped around pluripotency factors." *Nature*. doi: nature12420 [pii] 10.1038/nature12420.

Demare, L. E., J. Leng, J. Cotney, S. K. Reilly, J. Yin, R. Sarro, and J. P. Noonan. 2013. "The genomic landscape of cohesin-associated chromatin interactions." *Genome Res* 23 (8):1224-34. doi: gr.156570.113 [pii] 10.1101/gr.156570.113.

Denholtz, M., G. Bonora, C. Chronis, E. Splinter, W. de Laat, J. Ernst, M. Pellegrini, and K. Plath. 2013. "Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization." *Cell Stem Cell* 13 (5):602-16. doi: 10.1016/j.stem.2013.08.013.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485 (7398):376-80. doi: nature11082 [pii]

10.1038/nature11082.

Dowen, J. M., S. Bilodeau, D. A. Orlando, M. R. Hubner, B. J. Abraham, D. L. Spector, and R. A. Young. 2013. "Multiple Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements." *Stem Cell Reports* In Press.

Essafi, A., A. Webb, R. L. Berry, J. Slight, S. F. Burn, L. Spraggon, V. Velecela, O. M. Martinez-Estrada, J. H. Wiltshire, S. G. Roberts, D. Brownstein, J. A. Davies, N. D. Hastie, and P. Hohenstein. 2011. "A wt1-controlled chromatin switching mechanism underpins tissue-specific wnt4 activation and repression." *Dev Cell* 21 (3):559-74. doi: 10.1016/j.devcel.2011.07.014.

Felsenfeld, G., B. Burgess-Beusse, C. Farrell, M. Gaszner, R. Ghirlando, S. Huang, C. Jin, M. Litt, F. Magdinier, V. Mutskov, Y. Nakatani, H. Tagami, A. West, and T. Yusufzai. 2004. "Chromatin boundaries and chromatin domains." *Cold Spring Harb Symp Quant Biol* 69:245-50. doi: 10.1101/sqb.2004.69.245.

Filippova, D., R. Patro, G. Duggal, and C. Kingsford. 2014. "Identification of alternative topological domains in chromatin." *Algorithms Mol Biol* 9:14. doi: 10.1186/1748-7188-9-14.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. 2009. "An oestrogen-receptor-alpha-bound human chromatin interactome." *Nature* 462 (7269):58-64. doi: nature08497 [pii] 10.1038/nature08497.

Gibcus, J. H., and J. Dekker. 2013. "The hierarchy of the 3D genome." *Mol Cell* 49 (5):773-82. doi: S1097-2765(13)00139-1 [pii] 10.1016/j.molcel.2013.02.011.

Goh, Y., M. J. Fullwood, H. M. Poh, S. Q. Peh, C. T. Ong, J. Zhang, X. Ruan, and Y. Ruan. 2012. "Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation." *J Vis Exp* (62). doi: 10.3791/3770.

Gorkin, D. U., D. Leung, and B. Ren. 2014. "The 3D genome in transcriptional regulation and pluripotency." *Cell Stem Cell* 14 (6):762-75. doi: 10.1016/j.stem.2014.05.017.

Groschel, S., M. A. Sanders, R. Hoogenboezem, E. de Wit, B. A. Bouwman, C. Erpelinck, V. H. van der Velden, M. Havermans, R. Avellino, K. van Lom, E. J. Rombouts, M. van Duin, K. Dohner, H. B. Beverloo, J. E. Bradner, H. Dohner, B. Lowenberg, P. J. Valk, E. M. Bindels, W. de Laat, and R. Delwel. 2014. "A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia." *Cell* 157 (2):369-81. doi: 10.1016/j.cell.2014.02.019.

Guo, C., H. S. Yoon, A. Franklin, S. Jain, A. Ebert, H. L. Cheng, E. Hansen, O. Despo, C. Bossen, C. Vettermann, J. G. Bates, N. Richards, D. Myers, H. Patel, M. Gallagher, M. S. Schlissel, C. Murre, M. Busslinger, C. C. Giallourakis, and F. W. Alt. 2011. "CTCF-binding elements mediate control of V(D)J recombination." *Nature* 477 (7365):424-30. doi: 10.1038/nature10495.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. 2011. "CTCF-mediated functional chromatin interactome in pluripotent cells." *Nat Genet* 43 (7):630-8. doi: ng.857 [pii] 10.1038/ng.857.

Hawkins, R. D., G. C. Hon, C. Yang, J. E. Antosiewicz-Bourget, L. K. Lee, Q. M. Ngo, S. Klugman, K. A. Ching, L. E. Edsall, Z. Ye, S. Kuan, P. Yu, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, and B. Ren. 2011. "Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency." *Cell Res* 21 (10):1393-409. doi: 10.1038/cr.2011.146.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. 2013. "Super-enhancers in the control of cell identity and disease." *Cell* 155 (4):934-47. doi: 10.1016/j.cell.2013.09.053.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. 2010. "Mediator and cohesin connect gene expression and chromatin architecture." *Nature* 467 (7314):430-5. doi: 10.1038/nature09380.

Kaspi, H., E. Chapnik, M. Levy, G. Beck, E. Hornstein, and Y. Soen. 2013. "Brief report: miR-290-295 regulate embryonic stem cell differentiation propensities by repressing Pax6." *Stem Cells* 31 (10):2266-72. doi: 10.1002/stem.1465.

Kieffer-Kwon, K. R., Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grontved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan, and R. Casellas. 2013. "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation." *Cell* 155 (7):1507-20. doi: 10.1016/j.cell.2013.11.039.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren. 2007. "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." *Cell* 128 (6):1231-45. doi: S0092-8674(07)00205-X [pii] 10.1016/j.cell.2006.12.048.

Lee, T. I., R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. 2006. "Control of developmental regulators by Polycomb in human embryonic stem cells." *Cell* 125 (2):301-13. doi: S0092-8674(06)00384-9 [pii] 10.1016/j.cell.2006.02.043.

Lee, T. I., and R. A. Young. 2013. "Transcriptional regulation and its misregulation in disease." *Cell* 152 (6):1237-51. doi: 10.1016/j.cell.2013.02.014.

Lelli, K. M., M. Slattery, and R. S. Mann. 2012. "Disentangling the many layers of eukaryotic transcriptional regulation." *Annu Rev Genet* 46:43-68. doi: 10.1146/annurev-genet-110711-155437.

Li, G., M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H. S. Ooi, C. Tennakoon, C. L. Wei, Y. Ruan, and W. K. Sung. 2010. "ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing." *Genome Biol* 11 (2):R22. doi: 10.1186/gb-2010-11-2-r22.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. 2012. "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." *Cell* 148 (1-2):84-98. doi: S0092-8674(11)01517-0 [pii] 10.1016/j.cell.2011.12.014.

Margueron, R., and D. Reinberg. 2011. "The Polycomb complex PRC2 and its mark in life." *Nature* 469 (7330):343-9. doi: 10.1038/nature09784.

Merkenschlager, M., and D. T. Odom. 2013. "CTCF and cohesin: linking gene regulatory elements with their targets." *Cell* 152 (6):1285-97. doi: S0092-8674(13)00218-3 [pii] 10.1016/j.cell.2013.02.029.

Meuleman, W., D. Peric-Hupkes, J. Kind, J. B. Beaudry, L. Pagie, M. Kellis, M. Reinders, L. Wessels, and B. van Steensel. 2013. "Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence." *Genome Res* 23 (2):270-80. doi: 10.1101/gr.141028.112.

Naumova, N., M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker. 2013. "Organization of the mitotic chromosome." *Science* 342 (6161):948-53. doi: 10.1126/science.1236083.

Negre, N., J. Hennetin, L. V. Sun, S. Lavrov, M. Bellis, K. P. White, and G. Cavalli. 2006. "Chromosomal distribution of PcG proteins during Drosophila development." *PLoS Biol* 4 (6):e170. doi: 10.1371/journal.pbio.0040170.

Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard. 2012. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." *Nature* 485 (7398):381-5. doi: nature11049 [pii] 10.1038/nature11049.

Ong, C. T., and V. G. Corces. 2014. "CTCF: an architectural protein bridging genome topology and function." *Nat Rev Genet* 15 (4):234-46. doi: 10.1038/nrg3663.

Orkin, S. H., and K. Hochedlinger. 2011. "Chromatin connections to pluripotency and cellular reprogramming." *Cell* 145 (6):835-50. doi: 10.1016/j.cell.2011.05.019.

Parelho, V., S. Hadjur, M. Spivakov, M. Leleu, S. Sauer, H. C. Gregson, A. Jarmuz, C. Canzonetta, Z. Webster, T. Nesterova, B. S. Cobb, K. Yokomori, N. Dillon, L. Aragon, A. G. Fisher, and M. Merkenschlager. 2008. "Cohesins functionally associate with CTCF on mammalian chromosome arms." *Cell* 132 (3):422-33. doi: S0092-8674(08)00101-3 [pii] 10.1016/j.cell.2008.01.011.

Phillips, J. E., and V. G. Corces. 2009. "CTCF: master weaver of the genome." *Cell* 137 (7):1194-211. doi: S0092-8674(09)00699-0 [pii] 10.1016/j.cell.2009.06.001.

Phillips-Cremins, J. E., and V. G. Corces. 2013. "Chromatin insulators: linking genome organization to cellular function." *Mol Cell* 50 (4):461-74. doi: S1097-2765(13)00323-7 [pii] 10.1016/j.molcel.2013.04.018.

Phillips-Cremins, J. E., M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces. 2013. "Architectural protein subclasses shape 3D organization of genomes during lineage commitment." *Cell* 153 (6):1281-95. doi: S0092-8674(13)00529-1 [pii] 10.1016/j.cell.2013.04.053.

Roeder, R. G. 2005. "Transcriptional regulation and the role of diverse coactivators in animal cells." *FEBS Lett* 579 (4):909-15. doi: S0014-5793(04)01531-5 [pii] 10.1016/j.febslet.2004.12.007.

Rubio, E. D., D. J. Reiss, P. L. Welcsh, C. M. Disteche, G. N. Filippova, N. S. Baliga, R. Aebersold, J. A. Ranish, and A. Krumm. 2008. "CTCF physically links

cohesin to chromatin." *Proc Natl Acad Sci U S A* 105 (24):8309-14. doi: 0801273105 [pii] 10.1073/pnas.0801273105.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker. 2012. "The long-range interaction landscape of gene promoters." *Nature* 489 (7414):109-13. doi: nature11279 [pii] 10.1038/nature11279.

Schaaf, C. A., Z. Misulovin, M. Gause, A. Koenig, D. W. Gohara, A. Watson, and D. Dorsett. 2013. "Cohesin and polycomb proteins functionally interact to control transcription at silenced and active genes." *PLoS Genet* 9 (6):e1003560. doi: 10.1371/journal.pgen.1003560PGENETICS-D-13-00135 [pii].

Schwartz, Y. B., T. G. Kahn, D. A. Nix, X. Y. Li, R. Bourgon, M. Biggin, and V. Pirrotta. 2006. "Genome-wide analysis of Polycomb targets in Drosophila melanogaster." *Nat Genet* 38 (6):700-5. doi: 10.1038/ng1817.

Schwartz, Y. B., D. Linder-Basso, P. V. Kharchenko, M. Y. Tolstorukov, M. Kim, H. B. Li, A. A. Gorchakov, A. Minoda, G. Shanower, A. A. Alekseyenko, N. C. Riddle, Y. L. Jung, T. Gu, A. Plachetka, S. C. Elgin, M. I. Kuroda, P. J. Park, M. Savitsky, G. H. Karpen, and V. Pirrotta. 2012. "Nature and function of insulator protein binding sites in the Drosophila genome." *Genome Res* 22 (11):2188-98. doi: 10.1101/gr.138156.112.

Seitan, V. C., A. J. Faure, Y. Zhan, R. P. McCord, B. R. Lajoie, E. Ing-Simmons, B. Lenhard, L. Giorgetti, E. Heard, A. G. Fisher, P. Flicek, J. Dekker, and M. Merkenschlager. 2013. "Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments." *Genome Res* 23 (12):2066-77. doi: 10.1101/gr.161620.113.

Sexton, T., E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. 2012. "Three-dimensional folding and functional organization principles of the Drosophila genome." *Cell* 148 (3):458-72. doi: S0092-8674(12)00016-5 [pii] 10.1016/j.cell.2012.01.010.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. 2012. "A map of the cis-regulatory sequences in the mouse genome." *Nature* 488 (7409):116-20. doi: nature11243 [pii] 10.1038/nature11243.

Smallwood, A., and B. Ren. 2013. "Genome organization and long-range regulation of gene expression by enhancers." *Curr Opin Cell Biol* 25 (3):387-94. doi: S0955-0674(13)00028-8 [pii] 10.1016/j.ceb.2013.02.005.

Sofueva, S., E. Yaffe, W. C. Chan, D. Georgopoulou, M. Vietri Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur. 2013. "Cohesin-mediated interactions organize chromosomal domain architecture." *EMBO J* 32 (24):3119-29. doi: 10.1038/emboj.2013.237.

Soshnikova, N., T. Montavon, M. Leleu, N. Galjart, and D. Duboule. 2010. "Functional analysis of CTCF during mammalian limb development." *Dev Cell* 19 (6):819-30. doi: 10.1016/j.devcel.2010.11.009.

Spitz, F., and E. E. Furlong. 2012. "Transcription factors: from enhancer binding to developmental control." *Nat Rev Genet* 13 (9):613-26. doi: nrg3207 [pii] 10.1038/nrg3207.

Squazzo, S. L., H. O'Geen, V. M. Komashko, S. R. Krig, V. X. Jin, S. W. Jang, R. Margueron, D. Reinberg, R. Green, and P. J. Farnham. 2006. "Suz12 binds to silenced regions of the genome in a cell-type-specific manner." *Genome Res* 16 (7):890-900. doi: 10.1101/gr.5306606.

Tolhuis, B., E. de Wit, I. Muijrers, H. Teunissen, W. Talhout, B. van Steensel, and M. van Lohuizen. 2006. "Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster." *Nat Genet* 38 (6):694-9. doi: 10.1038/ng1792.

Valenzuela, L., and R. T. Kamakaka. 2006. "Chromatin insulators." *Annu Rev Genet* 40:107-38. doi: 10.1146/annurev.genet.39.073003.113546.

Van Bortle, K., E. Ramos, N. Takenaka, J. Yang, J. E. Wahi, and V. G. Corces. 2012. "Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains." *Genome Res* 22 (11):2176-87. doi: 10.1101/gr.136788.111.

Voss, T. C., R. L. Schiltz, M. H. Sung, P. M. Yen, J. A. Stamatoyannopoulos, S. C. Biddie, T. A. Johnson, T. B. Miranda, S. John, and G. L. Hager. 2011. "Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism." *Cell* 146 (4):544-54. doi: 10.1016/j.cell.2011.07.006.

Wang, H., C. Zang, L. Taing, K. L. Arnett, Y. J. Wong, W. S. Pear, S. C. Blacklow, X. S. Liu, and J. C. Aster. 2014. "NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers." *Proc Natl Acad Sci U S A* 111 (2):705-10. doi: 10.1073/pnas.1315023111.

Wen, B., H. Wu, Y. Shinkai, R. A. Irizarry, and A. P. Feinberg. 2009. "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." *Nat Genet* 41 (2):246-50. doi: 10.1038/ng.297.

Wendt, K. S., K. Yoshida, T. Itoh, M. Bando, B. Koch, E. Schirghuber, S. Tsutsumi, G. Nagae, K. Ishihara, T. Mishiro, K. Yahata, F. Imamoto, H. Aburatani, M. Nakao, N. Imamoto, K. Maeshima, K. Shirahige, and J. M. Peters. 2008. "Cohesin mediates transcriptional insulation by CCCTC-binding factor." *Nature* 451 (7180):796-801. doi: nature06634 [pii] 10.1038/nature06634.

Whyte, W. A., S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton, C. T. Foster, S. M. Cowley, and R. A. Young. 2012. "Enhancer decommissioning by

LSD1 during embryonic stem cell differentiation." *Nature* 482 (7384):221-5. doi: 10.1038/nature10805.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. 2013. "Master transcription factors and mediator establish super-enhancers at key cell identity genes." *Cell* 153 (2):307-19. doi: 10.1016/j.cell.2013.03.035.

Young, R. A. 2011. "Control of the embryonic stem cell state." *Cell* 144 (6):940-54. doi: 10.1016/j.cell.2011.01.032.

Zuin, J., J. R. Dixon, M. I. van der Reijden, Z. Ye, P. Kolovos, R. W. Brouwer, M. P. van de Corput, H. J. van de Werken, T. A. Knoch, W. F. van Ijcken, F. G. Grosveld, B. Ren, and K. S. Wendt. 2014. "Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells." *Proc Natl Acad Sci U S A* 111 (3):996-1001. doi: 10.1073/pnas.1317788111.

Chapter 4

Conclusions and future directions

Gene regulation is the process by which the genetic instructions stored in the DNA are selectively processed and interpreted by the cells. Understanding the regulation of gene expression is one of the fundamental goals of biological research. In the previous chapters, I presented two studies in which I designed computational methods to interrogate several types of large-scale genome-wide datasets that have provided new insights into the transcriptional and structural control of gene expression.

Chapter 2 described a study in which an expression-specificity approach was used to predict candidate master transcription factors in 100+ cell types. The study focused on the transcriptional control of retinal pigment epithelial (RPE) cells. These cells provide vital support to photoreceptor cells in the eye and their dysfunction is associated with the onset and progression of age-related macular degeneration (AMD). The predicted master transcription factors in RPE cells were used to guide the investigation of the transcriptional regulatory circuitry of these cells and to reprogram human fibroblasts into RPE-like cells. The RPE-like cells shared key features with RPE cells derived from healthy individuals, including morphology, gene expression and function. The identification of master transcription factors in the study should be useful for systematically mapping

regulatory circuitries and reprogramming cells for additional clinically relevant cell types.

Chapter 3 described the identification of a new type of chromosome structures called insulated neighborhoods that are functionally linked to gene control. Two key themes have emerged from studying these insulated neighborhood structures in mouse embryonic stem cells. First, the insulator protein CTCF confers insulator functions by forming CTCF-CTCF loops mediated by Cohesin. These loops likely create topological structures that constrain the physical contacts between cis-regulatory elements and their potential target genes. Second, insulated neighborhood structures formed by CTCF-CTCF/Cohesin loops are important for the proper expression of genes located either inside or immediately outside of these structures. More importantly, the key genes that define cell identity, including super-enhancer-driven genes and repressed genes encoding for regulators of other lineages, frequently occur within these insulated neighborhoods.

In the following sections in this chapter, I will describe how the computational approach to predict candidate master transcription factors can be adapted to next-generation sequencing-based gene expression datasets, and how the knowledge of master transcription factors can be applied to study the functional effects of genetic sequence variations. I will conclude by discussing how the integrity of insulated neighborhood structures may be disrupted in cancer genome.

Identification of master transcription factors using RNA-seq expression data

The expression-specificity approach described in chapter 2 uses an entropy-based measure to evaluate the relative expression levels and expression specificity of genes to predict candidate master transcription factors. The method quantifies the expression level of a transcript in a query cell type relative to the expression patterns of the transcript across a background dataset of diverse human cell and tissue types. The method requires the background dataset to be balanced to evenly represent the diversity of expression patterns of transcription factors. Expression datasets from the Affymetrix HG133 plus 2 platform were used to predict candidate master transcription factors because they represent the most comprehensive expression datasets available to date. The Human Body Index collection of microarray expression datasets (Gene Expression Omnibus, GSE7307) (Guo et al. 2013, Zhang et al. 2011) was used as the background dataset because the collection represents one of the largest and best curated repositories of expression datasets for human cell and tissue types at the time.

The expression-specificity approach should be applicable to various expression data types generated by other platforms. Among them, RNA-seq offers greater sensitivity and dynamic range in comparison to various types of expression microarrays. RNA-seq is a technique that profiles and quantifies transcriptome using next-generation high-throughput sequencing (Nagalakshmi et al. 2008, Wang et al. 2008, Zhang et al. 2012). Similar to hybridization-based expression microarrays, RNA-seq measures the abundance of cDNA fragments

representing the RNA population in the cells. These cDNA fragments are then sequenced by high-throughput sequencing technology and the resulting reads are aligned the genome to identify expressed transcripts and to determine their relative abundance. RNA-seq provides higher sensitivity of detection for genes expressed at both low and very high levels compared in comparison to expression microarrays.

More recently, transcriptome profiling using RNA-seq can be performed even at the single-cell level. This technology can reliably detect transcripts at the level of ~10 copies per cell (Grun, Kester, and van Oudenaarden 2014). As a result, single cell RNA-seq has become a powerful transcriptome discovery tool. It has already demonstrated its advantages in the studies of rare cell populations (e.g. stem cell population) or distinct cell types within a tissue (Jaitin et al. 2014, Tang et al. 2010, Zeisel et al. 2015). Since master transcription factors are expressed at relatively high levels, it should be feasible to predict candidate master transcription factors using single-cell RNA-seq data. The prediction using single-cell RNA-seq is likely to be more accurate than using expression profiles generated from large population of cells because the large population of cells may sometimes show strong cell-to-cell heterogeneity or represent different distinct cell types. In the vertebrate retina, for instance, there are more than ten different cell types. These cell types are often physically connected to each other and are difficult to dissect or separate experimentally. When applied to the mixture of different cell types in retina, single-cell RNA-seq should provide

insights into the transcriptome of these cells and help predict master transcription factors for potential cellular reprogramming applications in regenerative medicine.

Exploration of the effects of genetic sequence variations at the binding sites of master transcription factors

Genome-wide association studies (GWAS) have identified genomic loci where common genetic sequence variants (typically single nucleotide polymorphisms, SNPs) are statistically associated with complex human traits or diseases (Stranger, Stahl, and Raj 2011, Genomes Project et al. 2010). However, the GWAS studies usually do not distinguish functional genetic variants from the neutral ones that also show statistically significant association with traits or diseases. Although many early studies focused on the functional effects of genetic variants within coding regions or transcribed genes, increasing evidence suggests that genetic variants in non-coding regions also have strong influence in gene expression and organismal phenotype. It is estimated that over 70% of GWAS SNPs are located at non-coding regions of the genome. Typically, a large number of non-coding SNPs associated with a phenotype are distributed in proximity to each other in so-called linkage disequilibrium (LD). As a result, it remains challenging to identify the functional genetic sequence variants from the neutral variants in LD.

Identification of genetic sequences variants at the binding sites of master transcription factors can help identify functional genetic variants that are associated with complex human traits or diseases. For instance, a pioneering study demonstrated that the presence of SNPs in binding sites of transcription

factor nuclear factor kB (p65) leads to differences in transcription factor binding and gene expression between individuals (Kasowski et al. 2010). Owing to the recent advance of genome-wide mapping of cis-regulatory regions in human genome, it has also become apparent that the majority disease-associated sequence variation occurs in transcriptional regulatory regions in a cell-type-specific manner (Maurano et al. 2012, Hnisz et al. 2013). Furthermore, it has been proposed that functional variants tend to be located near the binding sites for master regulators of a given cell type (Farh et al. 2015). In combination of transcription factor consensus DNA binding motifs, the knowledge of master transcription factors in 100+ human tissue/cell types presented in chapter 2 should facilitate identification of functional genetic variants that are associated with complex human traits or diseases.

CTCF-CTCF DNA loops, insulated neighborhoods, and topological domains

Prior to the evidence presented in chapter 3, our understanding of topologically associated domains (TADs) is very coarse (Dixon et al. 2012, Nora et al. 2012). The boundaries of TADs are known to be enriched for a number of features, including CTCF binding, house-keeping genes, tRNAs, and SINE repeat elements (Dixon et al. 2012). However, it is not clear why the majority of DNA loop interactions are confined within TADs. The research in chapter 3 in this thesis has provided new insights into the molecular mechanisms that might explain the formation of TADs. There are ~8000 DNA Cohesin-associated loops interactions between CTCF bound regions that are not at enhancers and promoters in mouse embryonic stem cells. These CTCF-CTCF/cohesin loops

tend to be contained within TADs and to be enriched at the boundaries of TADs. This observation suggests that CTCF-CTCF/cohesin loops might be the molecular mechanism that creates the TAD boundaries.

A recent Hi-C study in human genome has lent support the idea (Rao et al. 2014). The study created a genome-wide interaction map at 1kb resolution and revealed that the genome is partitioned into ~9000 contact domains. These contact domains are averaging 185kb in size and frequently have CTCF occupied regions at their boundaries. Since the size of ~8000 CTCF-CTCF/Cohesin loops detected in mouse embryonic stem cells also averages around 200kb, these observations led me to postulate that the insulated neighborhood structures are actually the TAD domains. Although the previous studies about TADs prior to this study were able to identify ~2000 TAD structures in mammalian cell types, these numbers might have significantly underestimated the number of TADs because the resolution of hiC datasets was sufficiently high and computational algorithms only detected the strongest domain boundaries. Since Cohesin is associated with both CTCF-CTCF loops and enhancer-promoter loops, it will be important to study what extent Cohesin contribute the formation of TADs.

Mechanisms leading to the disruption of insulated neighborhood structures in cancer genome

Mutations or hypermethlation of CTCF consensus binding motifs can disrupt CTCF bindings and the associated insulated neighborhood structures.

CTCF is an 11-zinc finger protein that binds to a consensus DNA sequence CCCCTC. Its binding to the CTCF motif is negatively influenced by the DNA methylation (Engel et al. 2004, Wang et al. 2012). In cancer, somatic mutations (insertions, deletions, focal amplifications, and translocations) and aberrant hypermethylation of guanosine in CpG dinucleotide occur frequently (Hanahan and Weinberg 2011, Jones and Baylin 2002). It is conceivable that somatic mutations and hypermethylation disrupt CTCF binding sites result in loss of CTCF binding in cancer cells. If an insulated neighborhood structure is disrupted as a result of the loss of a CTCF binding site, the promoter of a gene contained within the insulated neighborhood might come into physical contact with distal active enhancers located outside of the insulated neighborhood. If this occurs to a proto-oncogene, the expression level of the proto-oncogene can be up-regulated contributing to oncogenesis.

Acknowledgements

I wish to thank members of the Young lab, especially Rick Young, and Alla Sigova, Lars Anders and Daniel Dadon for helpful comments during the preparation of this chapter.

References

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485 (7398):376-80. doi: nature11082 [pii] 10.1038/nature11082.

Engel, N., A. G. West, G. Felsenfeld, and M. S. Bartolomei. 2004. "Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations." *Nature Genetics* 36 (8):883-888. doi: DOI 10.1038/ng1399.

Farh, K. K., A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. 2015. "Genetic and epigenetic fine mapping of causal autoimmune disease variants." *Nature* 518 (7539):337-43. doi: 10.1038/nature13835.

Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. "A map of human genome variation from population-scale sequencing." *Nature* 467 (7319):1061-73. doi: 10.1038/nature09534.

Grun, D., L. Kester, and A. van Oudenaarden. 2014. "Validation of noise models for single-cell transcriptomics." *Nat Methods* 11 (6):637-40. doi: 10.1038/nmeth.2930.

Guo, J., M. Hammar, L. Oberg, S. S. Padmanabhuni, M. Bjareland, and D. Dalevi. 2013. "Combining evidence of preferential gene-tissue relationships from multiple sources." *PLoS One* 8 (8):e70568. doi: 10.1371/journal.pone.0070568.

Hanahan, D., and R. A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5):646-674. doi: DOI 10.1016/j.cell.2011.02.013.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. 2013. "Super-enhancers in the control of cell identity and disease." *Cell* 155 (4):934-47. doi: 10.1016/j.cell.2013.09.053.

Jaitin, D. A., E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. 2014. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types." *Science* 343 (6172):776-779. doi: DOI 10.1126/science.1247651.

Jones, P. A., and S. B. Baylin. 2002. "The fundamental role of epigenetic events in cancer." *Nature Reviews Genetics* 3 (6):415-428. doi: DOI 10.1038/nrg816.

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Y. Shi, A. E. Urban, M. Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder. 2010. "Variation in Transcription Factor Binding Among Humans." *Science* 328 (5975):232-235. doi: DOI 10.1126/science.1183621.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Z. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science* 337 (6099):1190-1195. doi: Doi 10.1126/Science.1222794.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. "The transcriptional landscape of the yeast genome defined by RNA sequencing." *Science* 320 (5881):1344-1349. doi: DOI 10.1126/science.1158441.

Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard. 2012. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." *Nature* 485 (7398):381-5. doi: nature11049 [pii] 10.1038/nature11049.

Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7):1665-1680. doi: Doi 10.1016/J.Cell.2014.11.021.

Stranger, B. E., E. A. Stahl, and T. Raj. 2011. "Progress and promise of genome-wide association studies for human complex trait genetics." *Genetics* 187 (2):367-83. doi: 10.1534/genetics.110.120907.

Tang, F. C., C. Barbacioru, S. Q. Bao, C. Lee, E. Nordman, X. H. Wang, K. Q. Lao, and M. A. Surani. 2010. "Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis." *Cell Stem Cell* 6 (5):468-478. doi: DOI 10.1016/j.stem.2010.03.015.

Wang, E. T., R. Sandberg, S. J. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. 2008. "Alternative isoform regulation in human tissue transcriptomes." *Nature* 456 (7221):470-476. doi: DOI 10.1038/nature07509.

Wang, H., M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers, and J. A. Stamatoyannopoulos. 2012. "Widespread plasticity in CTCF occupancy linked to DNA methylation." *Genome Res* 22 (9):1680-8. doi: 10.1101/gr.136101.111.

Zeisel, A., A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. Q. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. 2015. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq." *Science* 347 (6226):1138-1142. doi: DOI 10.1126/science.aaa1934.

Zhang, X., R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, H. Lv, and X. Li. 2011. "The expanded human disease network combining protein-protein interaction information." *Eur J Hum Genet* 19 (7):783-8. doi: 10.1038/ejhg.2011.30.

Zhang, Z., W. E. Theurkauf, Z. Weng, and P. D. Zamore. 2012. "Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection." *Silence* 3 (1):9. doi: 10.1186/1758-907X-3-9.

Appendix A

Supplementary material for chapter 2

SUPPLEMENTAL INFORMATION

SUPPLEMENTAL EXTENDED EXPERIMENTAL PROCEDURES

Identification of Candidate Master Transcription Factors

An entropy-based measure of Jensen-Shannon divergence (JSD) was adopted to evaluate the relative expression levels and expression specificity of transcription factors. The method quantified the expression level of a transcript in a query cell type relative to the expression patterns of the transcript across a background dataset of diverse human cell and tissue types. The major steps included collection of a background dataset, expression profile normalization, balancing of the background dataset, application of the JSD method, and integration of multiple datasets to generate a final ranking of transcription factors.

For the background dataset, 504 expression datasets, representing 106 cell and tissues types, were gathered primarily from the Human Body Index collection of expression datasets (Gene Expression Omnibus, GSE7307)(Guo et al. 2013, Zhang et al. 2011); the Human Body Index collection represents one of the largest and best curated repositories of expression datasets for human cell and tissue types. For additional cell and tissue types used as query datasets, publicly available expression datasets were used (Table S9).

All expression profiles used in this analysis were processed and normalized together to generate Affymetrix MAS5-normalized probe set values. CEL files were processed using the standard MAS5 normalization technique found in the affy package for the software program, R. The signals for multiple

individual probes assigned to a transcript were aggregated into a single probeset value using the standard probe assignment method ("hgu133plus2cdf").

The representation of cell and tissue types in the background dataset was balanced to evenly represent the diversity of expression patterns of transcription factors. If expression profiles from replicate samples or highly similar cell types are over-represented in the background expression dataset, the transcription factors that are highly specific to these cell types would be mistakenly considered as expressed in many different cell types. To construct a balanced background dataset, all profiles in the original background dataset were first clustered by similarity. Clusters of highly similar expression profiles were then identified, a single representative profile was chosen as the representative of the cluster, and other profiles in highly similar clusters were removed from the background dataset. For clustering, pair-wise comparisons were first performed on all expression profiles using Pearson correlation coefficients (PCCs). Hierarchal clustering then partitioned expression profiles into clusters based on the distance matrix derived from the PCCs. To choose a cutoff for partitioning expression profiles into clusters comprising highly similar expression profiles, the distribution of PCCs of expression profiles in the background dataset was empirically examined. The PCCs showed a bimodal distribution, suggesting there were two subpopulations of expression profiles, with the profiles of one group being more similar to each other. Examination of the profiles in the group with high PCCs indicated that many of the profiles were from redundant samples. This observation suggested that a cutoff separating the two subpopulations would be

generally useful for removing redundant profiles from the background dataset. This bimodal distribution was fitted with a mixture model with two Gaussian distributions to identify a cutoff value and a PCC of 0.9 was chosen to best separate the two subpopulations in the bimodal distribution. This cutoff was applied to identify clusters of similar profiles. Once clusters of similar profiles were identified, the medoid of a cluster was selected as the representative profile for that cluster of similar profiles. The expression profiles in the final, balanced background dataset are shown in Table S9.

Jensen-Shannon divergence (JSD), as described in (Fuglede 2004), was used to quantify the similarity between the observed pattern of transcription factor expression across cell types and the idealized pattern of a cell-type-specific master transcription factor across cell types. For each probeset that is mapped to a transcription factor, we created two same-sized, discrete probability vectors to represent the observed pattern and the ideal pattern. For the observed pattern, the vector was formed by values from the expression profiles of the query cell type and the balanced background dataset. The elements in this vector are divided by the sum so that the new normalized vector sums to 1. For the idealized pattern, the vector was formed by a value of 1 at the position equivalent to that of the query cell type and zeroes at all other positions. The distance metric between these two vectors was calculated using JSD and referred to as the cell-type-specificity score for the probeset. With this approach, the level of expression and the specificity of expression are incorporated into a single score, thus transcription factors scoring highly in either metric may score highly overall.

Where possible, multiple query datasets for a cell type of interest were used to identify candidate master transcription factors. The use of multiple query datasets theoretically helps identify the most robust candidate factors and should compensate to some degree for experimental and technical variability in gene expression experiments. One potential drawback is that datasets from different sources may purport to represent the same cells but may differ greatly due to differences in how the cells were obtained, heterogeneity of different cell populations or variations in growth conditions. If the differences between datasets are extreme, the use of multiple datasets may effectively cancel out relevant information. To compensate for this potential drawback, query datasets of the same cell types were compared by pair-wise Pearson correlation and datasets were grouped using hierarchical clustering. These subclusters can then be analyzed in a modular fashion, providing additional flexibility at this stage. Subclusters of datasets can be evaluated for suitability in inclusion, based on technical concerns. Subclusters of datasets may also reveal nuances of the underlying biology that may be instructive. For instance, subclusters that seem to represent different developmental stages of the same cell type may be separated at this stage, allowing for the selection of different sets of factors, biased by developmental stage. For this work, subclusters consisting of datasets that were largely dissimilar to other datasets (Pearson correlation coefficients less than 0.7 compared to other datasets) were removed from further consideration as we wished to provide a baseline set of candidate master transcription factors derived from the most representative, publicly available data.

To integrate information from multiple query datasets to yield a single ranking for a given cell or tissue type, rank product-based scores were next calculated for each probeset (Breitling et al. 2004). Only those query datasets that were retained after clustering as described above were included. Rank product-based scores tend to favor probesets that were ranked highly across multiple arrays and penalized probesets that scored highly in one or a few expression profiles. The main advantage of this rank product-based approach was that it favored consistency and did not require a "hard" cut-off when combining different datasets. The final ranked lists of candidate factors are provided in Table S1.

Microarray Expression Analysis for Knockdown Experiments

The raw data was obtained by using Affymetrix Gene Chip Operating Software using default settings. A Primeview CDF provided by Affymetrix was used to generate .CEL files. The CEL files were processed with the expresso command to convert the raw probe intensities to probeset expression values with MAS5 normalization using the standard tools available within the affy package in R. We used a loess regression (loess.normalize) from the affy package in R to renormalize the probe values using only the probes mapped to ribosomal genes to fit the loess. For genes with multiple probesets, the probeset with the maximum signal across experiments was selected for further analysis. Differential gene expression was determined using moderated t-statistic in the "limma" package (http://bioinf.wehi.edu.au/limma/) from Bioconductor

(www.bioconductor.org) (Smyth 2004). Two independent hairpins were treated as replicates and compared to the two control hairpins. A gene was considered differentially expressed if it met the following criteria: 1) absolute log2 fold-change ≥1 between the mean expression of the two control shRNAs and the mean expression of the two target shRNAs, 2) adjusted p-value ≤ 0.1 by a moderated t-test within the limma package with BH multiple hypothesis testing correction. Expression change of all RefSeq genes after shRNA knockdown in RPE cells is shown in Table S4. Raw data and processed gene expression tables can be found online associated with the raw and processed sequencing and microarray data were deposited in GEO under accession numbers GSE60024 and GSE64264 (http://www.ncbi.nlm.nih.gov/geo/).

Determining Enriched GO terms

The nature of differentially expressed genes was examined using GO analysis. Enriched Gene Ontology classification terms were identified using GO Term finder (http://go.princeton.edu/cgi-bin/GOTermFinder). The differentially up- and down-regulated genes from different candidate master transcription factor knockdown experiments were pooled together and used as inputs. The default settings of hypergeometric test with multiple hypothesis Bonferroni correction (adjusted p-Values of 0.01) was used.

Gene Set Enrichment Analysis (GSEA)

GSEA (Broad Institute, http://www.broadinstitute.org/gsea/) was performed for differentiated expressed genes pooled from different candidate master transcription factor knockdown experiments. The differentially expressed genes were pre-ranked by the average fold change (log2) in cells harboring transcription factor knockdown constructs relative to cells harboring the non-targeting shRNA control. The published RPE signature genes (Strunnikova et al. 2010) were used as the gene set for enrichment analysis.

Chromatin Immunoprecipitation (ChIP)

ChIP protocols have previously been described in detail (Lee, Johnstone, and Young 2006). RPE cells were grown to passage 4 and crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 12 minutes at room temperature. Cells were rinsed twice with 1x PBS, pelleted by centrifugation and flash frozen in liquid nitrogen and stored at –80°C. Cell pellets were resuspended, lysed and sonicated to solubilize and shear crosslinked DNA. We used a Bioruptor (Diagenode) and sonicated at medium power for 10 x 30 second pulses (30 second pause between pulses). Samples were kept on ice at all times. The resulting input material was incubated overnight at 4°C with 20 µl of Dynal Protein G magnetic beads (Life Technologies, cat. #10004D) that had been pre-incubated with 5 µg of the appropriate antibody. The immunoprecipitation was allowed to proceed overnight at 4°C. For MITF, OTX2, PAX6, ZNF92, LHX2 immunoprecipitations, beads were washed twice with 20mM Tris-HCl pH8.0, 150 mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100,

once with 20mM Tris-HCl pH8.0, 500mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100, once with 10mM Tris-HCl pH8.0, 250nM LiCl, 2mM EDTA, 1% NP40 and once with TE containing 50 mM NaCl. For RNA Pol II and H3K27Ac immunoprecipitations, sodium deoxycholate (0.1% final concentration) was added to all washes except the final TE wash. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing and crosslinking was reversed by incubation at 65°C for eight hours. Input material DNA (reserved from the sonication step) was also treated for crosslink reversal. Immunoprecipitated DNA and input material DNA were then purified by treatment with RNAse A, proteinase K and phenol:chloroform:isoamyl alcohol extraction. The antibodies used for ChIP analysis are listed in table S5.

Illumina Sequencing and Library Generation

Purified ChIP DNA was used to prepare Illumina multiplexed sequencing libraries. Libraries for Illumina sequencing were prepared following the Illumina TruSeq DNA Sample Preparation v2 kit protocol with the following exceptions. After end-repair and A-tailing, immunoprecipitated DNA (~10-50 ng) or input DNA (50 ng) was ligated with a 1:50 dilution of Illumina Adaptor Oligo Mix assigning one of 24 unique index primer sets in the kit to each sample. Following ligation, libraries were amplified by 18 cycles of PCR using the HiFi NGS Library Amplification kit from KAPA Biosystems. Amplified libraries were then size-selected using a 2% gel cassette in the Pippin Prep system from Sage Science set to capture fragments between 200 and 400 bp. Libraries were quantified by

qPCR using the KAPA Biosystems Illumina Library Quantification kit according to kit protocols. Libraries with distinct TruSeq index primers were multiplexed by mixing at equimolar ratios and running together in a lane on the Illumina HiSeq 2000 for 40 bases in single read mode.

ChIP-Seq Data Analysis

All ChIP-Seq datasets were aligned to build version NCBI37/HG19 of the human genome using Bowtie (version 0.12.9) (Langmead et al. 2009) with the following parameters: –n2, -e70, -m2, -k2, --best. We used the MACS version 1.4.1 (Model based analysis of ChIP-Seq) (Zhang et al. 2008) peak finding algorithm to identify regions of ChIP-Seq enrichment over background. A p-value threshold of enrichment of 1e-7 was used for all datasets with parameter --no-model, -dup=2. Approximately 15,200, 13,700, 9,400, 3,300, 12,500, regions were identified for LHX2, OTX2, PAX6, MITF, ZNF92, respectively. Wiggle files for gene tracks were created using MACS with options –w –S –space=50 to count reads in 50bp bins. They were normalized to the total number (in millions) of mapped reads producing the final tracks in units of reads per million mapped reads per bp (rpm/bp).

Assigning Genes to Transcription Factor Binding Sites

All analyses were performed using RefSeq (GRCh37/hg19) human gene annotations. A gene was defined as transcribed if an enriched region for H3K27ac or RNA Pol II was located at the TSS. Active genes were assigned to

transcription binding sites using the following method. Using a simple proximity rule, for each ChIP enriched region, the nearest TSS of an active gene was assigned to the region. Since promoters and distal elements can engage in looping interactions beyond the nearest genes (Thurman et al. 2012b), additional genes were assigned to ChIP enriched regions by using the distal DHS-to-promoter connection maps from a recent large-scale ENCODE study of promoters and their co-regulated distal DHS in 79 human cell types (Thurman et al. 2012a). For each ChIP enriched region overlapping with a distal DHS in the distal DHS-to-promoter connection map, the genes from the DHS-to-promoter pair were assigned to the region.

Definition of Active Enhancers

Active enhancers were defined as regions showing enrichment for H3K27Ac outside of promoters (greater than 2.5kb away from any TSS). H3K27Ac is a histone modification associated with active enhancers (Creyghton et al. 2010, Rada-Iglesias et al. 2011).

Identifying Super-Enhancers

The identification of super-enhancers previously been described in detail (Hnisz et al. 2013). Briefly, H3K27ac peaks were used to identify constituent enhancers. These were stitched if within 12.5kb, and peaks fully contained within +/- 2kb from a TSS were excluded from stitching. H3K27ac signal (less input control) was used to rank enhancers by their enrichment. 670 super-enhancers

were separated from typical enhancers as previously described (Whyte et al. 2013, Loven et al. 2013). Super-enhancers were assigned to active genes using the ROSE software package (younglab.wi.mit.edu/super_enhancer_code.html). The super-enhancers and their target genes are listed in Table S6.

Principal Component Analysis and Differential Expression Analysis for iRPE

All expression datasets used for this analysis were processed together to generate Affymetrix MAS5-normalized probe set values. We processed all CEL files by using the probe definition ("hgu133plus2cdf") and the standard MAS5 normalization technique within the affy package in R to get probe set expression values. The probesets of the same gene were next collapsed into a single value to represent the gene by taking the values of the probeset with the maximum signal across experiments.

The top 25% genes with the largest coefficient of variation across all expression profiles were used for Principal Component Analysis (PCA). PCA was done using R and the package MADE4 (Culhane et al. 2005). Previously published microarray data used in PCA analysis is listed in Table S9.

Differential gene expression between human foreskin fibroblasts (HFF) and retinal pigment epithelial (RPE) cells was determined using moderated t-statistic in the "limma" package (http://bioinf.wehi.edu.au/limma/) from Bioconductor (www.bioconductor.org) (Smyth 2004). The differentially expressed genes were required to have absolute value of log2 fold-change ≥1 between the

mean expression of HFFs and the mean expression RPEs, and FDR-adjusted p-value ≤ 0.01. The heat map in Figure 4F shows the fold change (log2) of the differentially expressed genes relative to the mean expression of HFF.

SUPPLEMENTAL REFERENCES

Breitling, R., P. Armengaud, A. Amtmann, and P. Herzyk. 2004. "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." *FEBS Lett* 573 (1-3):83-92. doi: 10.1016/j.febslet.2004.07.055.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 2010. "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50):21931-21936. doi: Doi 10.1073/Pnas.1016071107.

Culhane, A. C., J. Thioulouse, G. Perriere, and D. G. Higgins. 2005. "MADE4: an R package for multivariate analysis of gene expression data." *Bioinformatics* 21 (11):2789-90. doi: 10.1093/bioinformatics/bti394.

Fuglede, B., and Topsoe, F. 2004. "Jensen-Shannon Divergence and Hilbert space embedding." *Information theory* 31.

Guo, J., M. Hammar, L. Oberg, S. S. Padmanabhuni, M. Bjareland, and D. Dalevi. 2013. "Combining evidence of preferential gene-tissue relationships from multiple sources." *PLoS One* 8 (8):e70568. doi: 10.1371/journal.pone.0070568.

Hnisz, D., B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. 2013. "Super-enhancers in the control of cell identity and disease." *Cell* 155 (4):934-47. doi: 10.1016/j.cell.2013.09.053.

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

Lee, T. I., S. E. Johnstone, and R. A. Young. 2006. "Chromatin immunoprecipitation and microarray-based analysis of protein location." *Nat Protoc* 1 (2):729-48. doi: nprot.2006.98 [pii] 10.1038/nprot.2006.98.

Loven, J., H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young. 2013. "Selective inhibition of tumor

oncogenes by disruption of super-enhancers." *Cell* 153 (2):320-34. doi: S0092-8674(13)00393-0 [pii] 10.1016/j.cell.2013.03.036.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. "A unique chromatin signature uncovers early developmental enhancers in humans." *Nature* 470 (7333):279-283. doi: Doi 10.1038/Nature09692.

Smyth, G. K. 2004. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol* 3:Article3. doi: 10.2202/1544-6115.1027.

Strunnikova, N. V., A. Maminishkis, J. J. Barb, F. Wang, C. Zhi, Y. Sergeev, W. Chen, A. O. Edwards, D. Stambolian, G. Abecasis, A. Swaroop, P. J. Munson, and S. S. Miller. 2010. "Transcriptome analysis and molecular signature of human retinal pigment epithelium." *Hum Mol Genet* 19 (12):2468-86. doi: ddq129 [pii] 10.1093/hmg/ddq129.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. 2012a. "The accessible chromatin landscape of the human genome." *Nature* 489 (7414):75-82. doi: 10.1038/nature11232.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Z. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Y. Song, S. Vong, M. Weaver, Y. Q. Yan, Z. C. Zhang, Z. Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. 2012b. "The accessible chromatin landscape of the human genome." *Nature* 489 (7414):75-82. doi: Doi 10.1038/Nature11232.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. 2013. "Master transcription factors and mediator establish super-enhancers at key cell identity genes." *Cell* 153 (2):307-19. doi: S0092-8674(13)00392-9 [pii] 10.1016/j.cell.2013.03.035.

Zhang, X., R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, H. Lv, and X. Li. 2011. "The expanded human disease network combining protein-protein interaction information." *Eur J Hum Genet* 19 (7):783-8. doi: 10.1038/ejhg.2011.30.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol* 9 (9):R137. doi: gb-2008-9-9-r137 [pii]10.1186/gb-2008-9-9-r137.

Appendix B

Supplementary material for chapter 2

SUPPLEMENTAL INFORMATION

Supplemental Data

SUPPLEMENTAL DATA

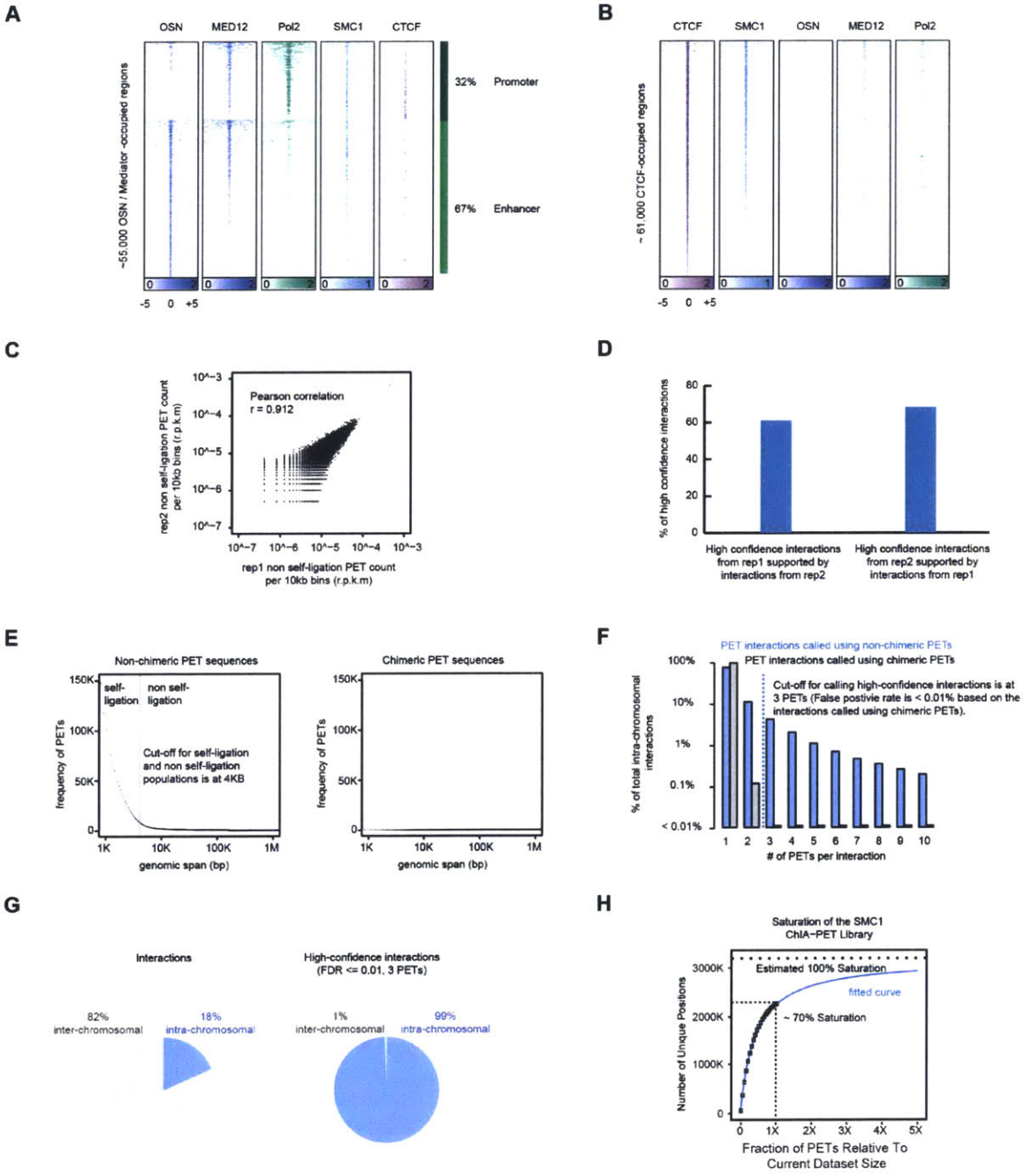


Figure S1. PET quality assessment and interactions. Related to Figure 1.

A) Heatmap representation of ESC ChIP-seq data for the combination of the master transcription factors OCT4, SOX2 and NANOG (OSN), MED12, RNA

polymerase II (Pol2), CTCF, and SMC1 at promoters and enhancers in ESCs. Read density is displayed within a 10kb window and color scale intensities are shown in rpm/bp.

B) Heatmap representation of ESC ChIP-seq data for the combination of the master transcription factors OCT4, SOX2 and NANOG (OSN), MED12, RNA polymerase II (Pol2), CTCF, and SMC1 at CTCF-bound sites in ESCs. Read density is displayed within a 10kb window and color scale intensities are shown in rpm/bp.

C) Scatter plot showing the number of non self-ligation PETs per 10kb in replicates in reads per million mapped reads per kilobase.

D) Bar graph showing the percentage of high confidence interactions from one replicate of the SMC1 ChIA-PET being supported by interactions in the other replicate.

E) *Left,* scatter plot showing the frequency of non-chimeric PETs with homodimeric linkers against PET genomic span in increments of 100 bp. The curve suggests a distance cut-off at ~4 kb, below which the PET sequences may originate from self-ligation of DNA ends from a single chromatin fragment in the ChIA-PET protocol. *Right,* scatter plot showing chimeric PET frequencies with heterodimeric linkers against PET genomic span in increments of 100 bp, suggesting chimeric PETs were more uniformly distributed across different genomic spans.

F) Bar graph showing the percentage of interactions called by requiring different numbers of chimeric and non-chimeric PETs. All PET interactions called using

chimeric PETs that are supported by at least 3 PETs have a false discovery rate <0.01%.

G) Diagram showing the frequency of intrachromosomal and interchromosomal interactions in the interaction (left) and high confidence interaction dataset (right).

H) Saturation analysis of the SMC1 ChIA-PET dataset. Subsampling of various fractions of PETs within the merged ChIA-PET dataset was performed, and the number of unique genomic positions of intrachromosomal PETs beyond the self-ligation distance cutoff of 4 kb was plotted. The solid line depicting the non-linear least-squares regression fitting of the data to the Michaelis-Menten model suggests that we have sampled approximately 70% of the available intrachromosomal PETs beyond 4 kb in the current library. The dashed line indicates the estimated 100% saturation.

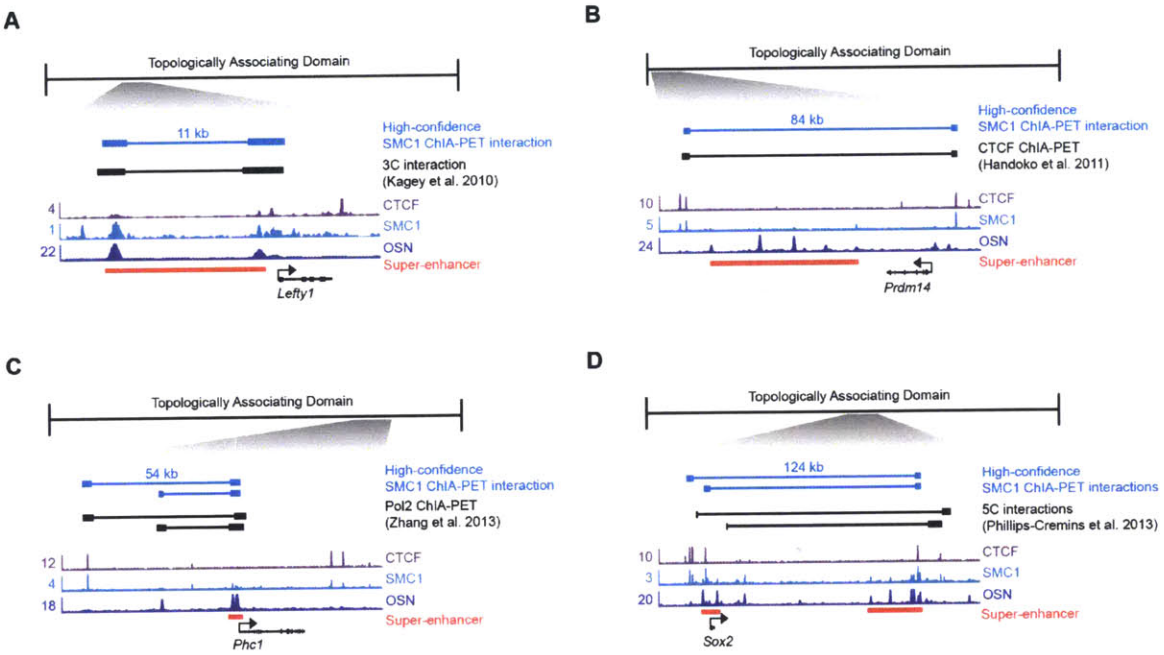


Figure S2. High-confidence SMC1 ChIA-PET interactions are consistent with previously identified interactions. Related to Figure 1.

High-confidence SMC1-ChIA-PET interactions are depicted as blue lines. Interactions from other published datasets are depicted as black lines. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and OCT4, SOX2, and NANOG (OSN) are shown at the indicated loci in ESCs.

A) A high-confidence SMC1-ChIA-PET interaction is supported by 3C from (Kagey et al. 2010). Genomic coordinates for the *Lefty1* TAD are chr1:182,760,000-183,160,000. Genomic coordinates for the *Lefty1* ChIP-Seq binding profiles are chr1:182,851,700-182,871,500.

B) A high-confidence SMC1-ChIA-PET interaction is supported by a CTCF ChIA-PET PET from (Handoko et al. 2011). Genomic coordinates for the *Prdm14* TAD are chr1:13,040,000-13,680,000. Genomic coordinates for the *Prdm14* ChIP-Seq binding profiles are chr1:13,034,300-13,131,900.

C) A high-confidence SMC1-ChIA-PET interaction is supported by a Pol2 ChIA-PET PET in (Zhang et al. 2013). Genomic coordinates for the *Phc1* TAD are chr6:121,160,000-122,600,000. Genomic coordinates for the *Phc1* ChIP-Seq binding profiles are chr6:122,241,500-122,350,700.

D) A high-confidence SMC1-ChIA-PET interaction is supported by 5C in (Phillips-Cremins et al. 2013). Genomic coordinates for the *Sox2* TAD are chr3:33,680,000-35,520,000. Genomic coordinates for the *Sox2* ChIP-Seq binding profiles are chr3:34,522,100-34,691,600.

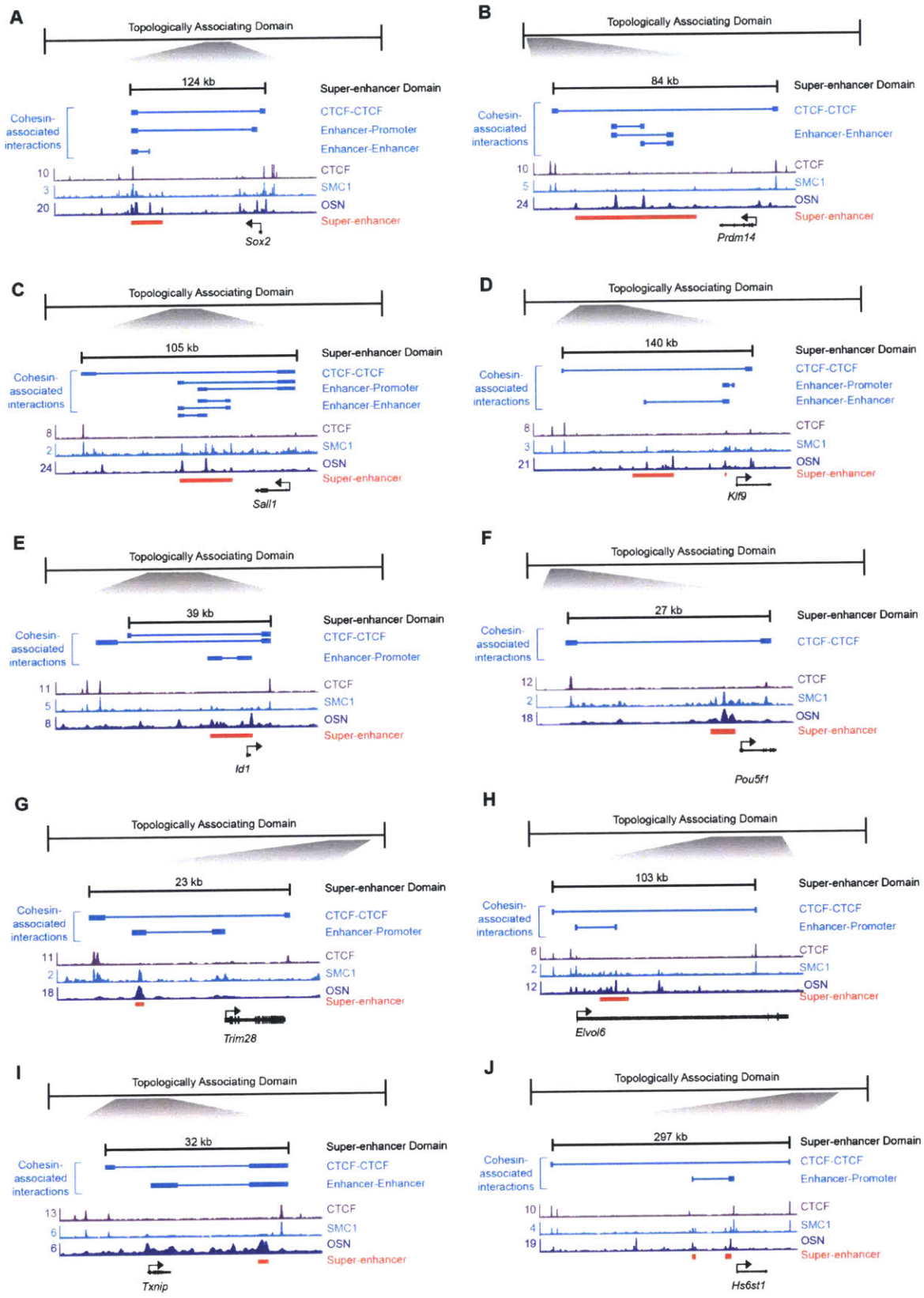


Figure S3. Super-enhancer Domains. Related to Figure 3.

Active cell identity genes reside in Super-enhancer Domains (SD). Shown are example SDs within Topologically Associating Domains (TADs) in ESCs. High-confidence SMC1 ChIA-PET interactions are depicted as blue lines. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and the master transcription factors OCT4, SOX2, and NANOG (OSN) are shown at the example SDs in ESCs. Super-enhancer regions are indicated by a red bar.

A) Genomic coordinates for the *Sox2* TAD are chr3:35,520,000-33,680,000. Genomic coordinates for the *Sox2* binding profiles are chr3:34,724,900-34,502,100.

B) Genomic coordinates for the *Prdm14* TAD are chr1:13,040,000-13,680,000. Genomic coordinates for the *Prdm14* binding profiles are chr1:13,034,300-13,131,900.

C) Genomic coordinates for the *Sall1* TAD are chr8:90,920,000-92,360,000. Genomic coordinates for the *Sall1* binding profiles are chr8:91,455,200-91,581,300.

D) Genomic coordinates for the *Klf9* TAD are chr19:22,920,000-24,360,000. Genomic coordinates for the *Klf9* and binding profiles are chr19:23,068,300-23,273,400.

E) Genomic coordinates for the *Id1* TAD are chr2:152,440,000-152,680,000. Genomic coordinates for the *Id1* binding profiles are chr2:152,511,000-152,581,000.

F) Genomic coordinates for the *Pou5f1* TAD are chr17: 35,600,000-36,080,000. Genomic coordinates for the *Pou5f1* binding profiles are chr17:35,617,300-

35,649,800.

G) Genomic coordinates for the *Trim28* TAD are chr7:13,000,000-13,640,000. Genomic coordinates for the *Trim28* binding profiles are chr7:13,590,396-13,620,304.

H) Genomic coordinates for the *Elovl6* TAD are chr3:128,920,000-129,480,000. Genomic coordinates for the *Elovl6* binding profiles are chr3:129,217,096-129,348,924.

I) Genomic coordinates for the *Txnip* TAD are chr3:96,320,000-96,520,000. Genomic coordinates for the *Txnip* binding profiles are chr3:96,347,300-96,391,100.

J) Genomic coordinates for *Hs6st1* TAD are chr1:34,520,000-36,360,000. Genomic coordinates for *Hs6st1* binding profiles are chr1:35,883,900-36,200,400.

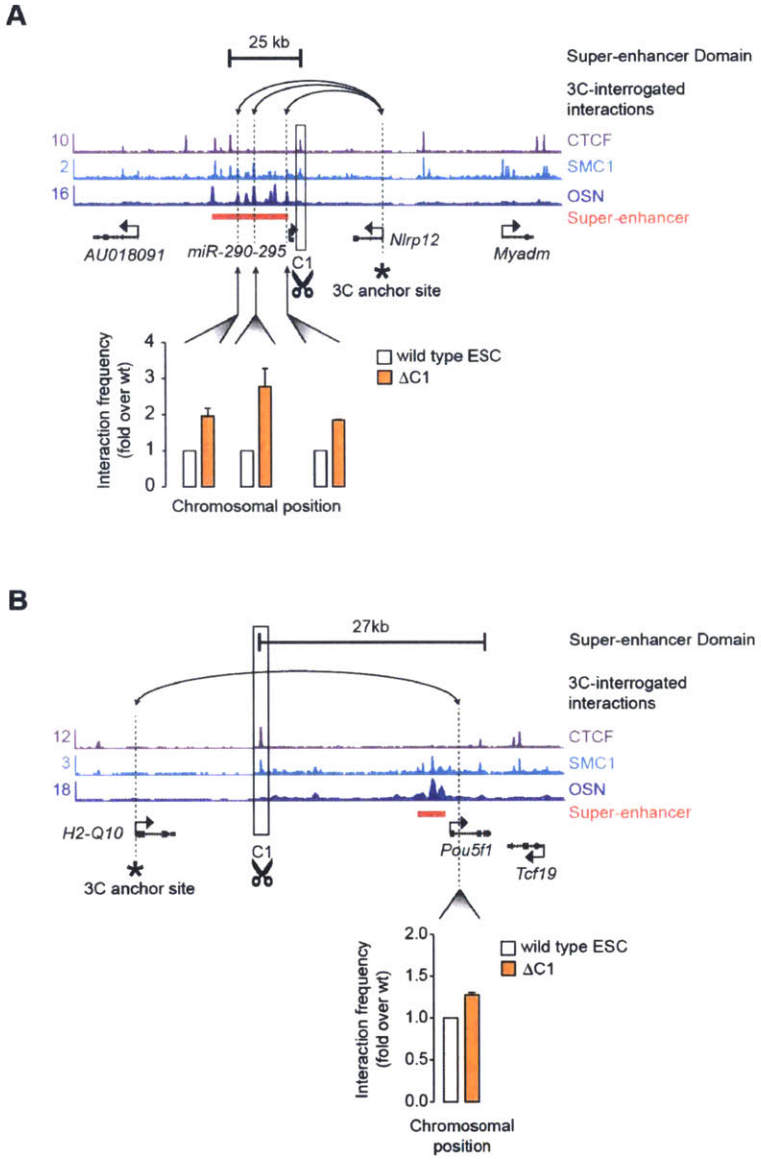

A
25 kb
Super-enhancer Domain
3C-interrogated interactions
10
2
16
CTCF
SMC1
OSN
Super-enhancer
AU018091
miR-290-295
C1
Nlrp12
Myadm
3C anchor site
wild type ESC
ΔC1
Interaction frequency (fold over wt)
Chromosomal position
B
27kb
Super-enhancer Domain
3C-interrogated interactions
12
3
18
CTCF
SMC1
OSN
Super-enhancer
H2-Q10
C1
Pou5f1
Tcf19
3C anchor site
wild type ESC
ΔC1
Interaction frequency (fold over wt)
Chromosomal position
C
wild type ESC
ΔC1 (miR-290-295)
Expression level (% to wild type)
Pou5f1
Sox2
Nanog
Pax6
Gata6
Sox17
Pluripotency marker
Ectoderm marker
Endoderm marker

Figure S4. Super-enhancer Domain functions. Related to Figure 4.

A) Quantitative 3C analysis at the *miR-290-295* locus. The super-enhancer domain is indicated as a black bar. The deleted CTCF site is highlighted with a box. Arrows indicate the chromosomal positions between which the interaction frequency was assayed. Asterisk indicates the 3C anchor site. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and the master transcription factors OCT4, SOX2, and NANOG (OSN) are also shown. The super-enhancer is indicated as a red bar. The interaction frequencies between the indicated chromosomal positions and the 3C anchor sites are displayed as a bar chart on the bottom panel. qPCR reactions were run in duplicates, and values are normalized against the mean interaction frequency in wild type cells. ($P<0.05$ for all three regions; Student´s t-test.)

B) Quantitative 3C analysis at the *Pou5f1* locus. The super-enhancer domain is indicated as a black bar. The deleted CTCF site is highlighted with a box. Arrow indicate the chromosomal positions between which the interaction frequency was assayed. Asterisk indicates the 3C anchor site. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and the master transcription factors OCT4, SOX2, and NANOG (OSN) are also shown. The super-enhancer is indicated as a red bar. The interaction frequencies between the indicated chromosomal positions and the 3C anchor sites are displayed as a bar chart on the bottom panel. qPCR reactions were run in duplicates, and values are normalized against the mean interaction frequency in wild type cells. ($P<0.05$; Student´s t-test.)

C) Expression level of the indicated germ layer markers in wild type cells and a cell line where the SD boundary CTCF site was deleted at the *miR-290-295* locus. Gene expression was measured by qRT-PCR. Gene expression was assayed in triplicate reactions in at least two biological replicate samples (P-value < 0.003, *PAX6, GATA6 and Sox17* in wild-type vs. CTCF site-deleted). P-value was calculated using the Student's t-test.

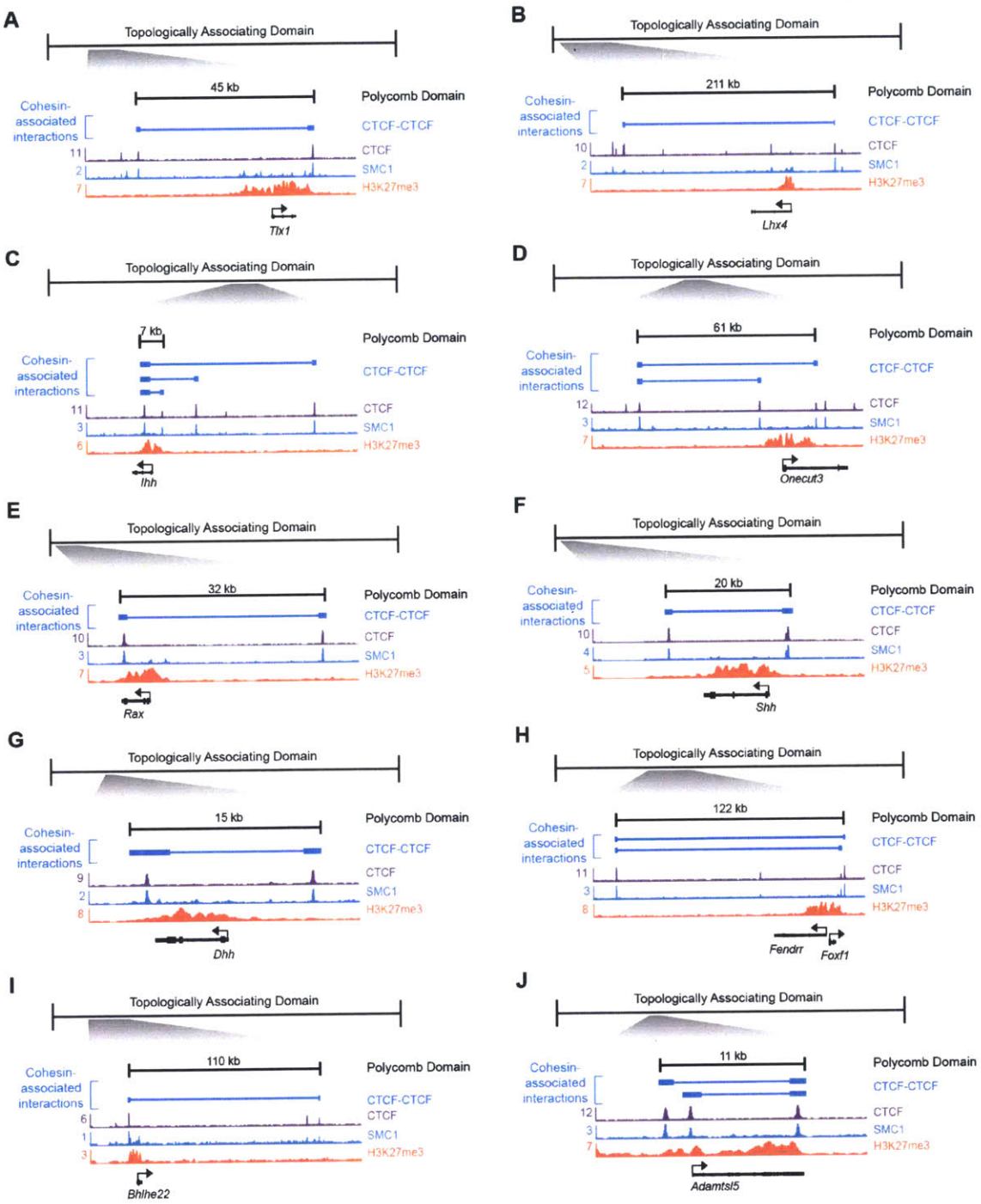


Figure S5. Polycomb Domain interactions. Related to Figure 5.

Repressed developmental lineage genes reside in chromosome structures termed Polycomb Domains (PD). Example PDs within Topologically Associating

Domains (TADs) are shown with high-confidence PET interactions depicted by blue lines. ChIP-Seq binding profiles (reads per million per base pair) for CTCF, cohesin (SMC1), and H3K27me3 are shown at the example PDs in ESCs.

A) Genomic coordinates for the *Tlx1* TAD are chr19:45,120,000-45,840,000. Genomic coordinates for the *Tlx1* binding profiles are chr19:45,178,400-45,246,700.

B) Genomic coordinates for the *Lhx4* TAD are chr1:157,400,000-158,640,000. Genomic coordinates for the *Lhx4* binding profiles are chr1:157,392,000-157,657,700.

C) Genomic coordinates for the *Ihh* TAD are chr1:74,240,000-75,600,000. Genomic coordinates for the *Ihh* binding profiles are chr1:74,978,200-75,060,400.

D) Genomic coordinates for the *Onecut3* TAD are chr10:79,200,001-81,040,000. Genomic coordinates for the *Onecut3* binding profiles are chr10:79,892,959-79,985,160.

E) Genomic coordinates for the *Rax* TAD are chr18:66,080,001-66,680,000. Genomic coordinates for the *Rax* binding profiles are chr18:66,089,130-66,130,404.

F) Genomic coordinates for the *Shh* TAD are chr5:28,760,001-29,680,000. Genomic coordinates for the *Shh* binding profiles are chr5:28,766,181-28,808,422.

G) Genomic coordinates for the *Dhh* TAD are chr15:98,360,001-100,560,000. Genomic coordinates for the *Dhh* binding profiles are chr15:98,718,426-98,738,916.

H) Genomic coordinates for the *Fendrr/Foxf1* TAD are chr8:123,160,001-124,360,000. Genomic coordinates for the *Fendrr/Foxf1* binding profiles are chr8:123,482,102-123,627,553.

I) Genomic coordinates for the *Bhlhe22* TAD are chr3:17,800,001-19,120,000. Genomic coordinates for the *Bhlhe22* binding profiles are chr3:17,927,749-18,082,958.

J) Genomic coordinates for the *Adamtls5* TAD are chr10:79,200,001-81,040,000. Genomic coordinates for the *Adamtls5* binding profiles are chr10:79,797,646-79,818,602.

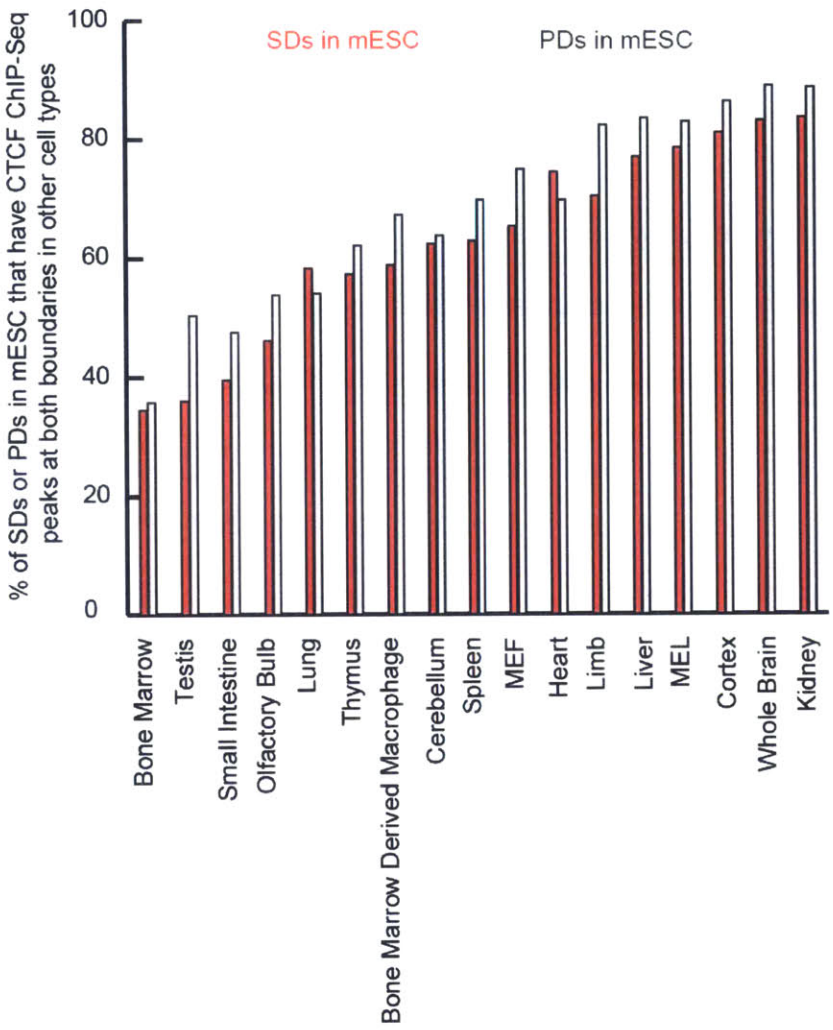


Figure S6. SD and PD boundary sites are constitutively occupied by CTCF across multiple cell types. Related to Figure 6.

The proportions of SDs and PDs identified in ESCs for which CTCF ChIP-seq peaks at both boundaries are observed in other mouse cell types. Occupancy of CTCF peaks across the cell types was determined from publicly available CTCF ChIP-seq data (Shen et al. 2012). MEF cells are murine embryonic fibroblasts and MEL cells are murine erythroleukemia cells.

SUPPLEMENTAL EXTENDED EXPERIMENTAL PROCEDURES

Cell Culture

V6.5 murine ESCs were grown on irradiated murine embryonic fibroblasts (MEFs). Cells were grown under standard ESC conditions as described previously (Whyte et al. 2012). Cells were grown on 0.2% gelatinized (Sigma, G1890) tissue culture plates in ESC media; DMEM-KO (Invitrogen, 10829-018) supplemented with 15% fetal bovine serum (Hyclone, characterized SH3007103), 1,000 U/ml LIF (ESGRO, ESG1106), 100 mM nonessential amino acids (Invitrogen, 11140-050), 2 mM L-glutamine (Invitrogen, 25030-081), 100 U/ml penicillin, 100 mg/ml streptomycin (Invitrogen, 15140-122), and 8 nl/ml of 2-mercaptoethanol (Sigma, M7522).

ChIA-PET Library Construction

ChIA-PET was performed as previously described (Fullwood et al. 2009, Goh et al. 2012, Li et al. 2012, Chepelev et al. 2012). Briefly, ES cells (up to $1x10^8$ cells) were treated with 1% formaldehyde at room temperature for 20 min and then neutralized using 0.2M glycine. The crosslinked chromatin was fragmented by sonication to size lengths of 300-700 bp. The anti-SMC1 antibody (Bethyl, A300-055A) was used to enrich SMC1-bound chromatin fragments. A portion of ChIP DNA was eluted from antibody-coated beads for concentration quantification and for enrichment analysis using quantitative PCR. For ChIA-PET library construction ChIP DNA fragments were end-repaired using T4 DNA

polymerase (NEB). ChIP DNA fragments were divided into two aliquots and either linker A or linker B was ligated to the fragment ends. The two linkers differ by two nucleotides which are used as a nucleotide barcode (Linker A with CG; Linker B with AT) (Table S1). After linker ligation, the two samples were combined and prepared for proximity ligation by diluting in a 20 ml volume to minimize ligations between different DNA-protein complexes. The proximity ligation reaction was performed with T4 DNA ligase (Fermentas) and incubated without rocking at 22 degrees Celsius for 20 hours. During the proximity ligation DNA fragments with the same linker sequence were ligated within the same chromatin complex, which generated the ligation products with homodimeric linker composition. However, chimeric ligations between DNA fragments from different chromatin complexes could also occur, thus producing ligation products with heterodimeric linker composition. These heterodimeric linker products were used to assess the frequency of nonspecific ligations and were then removed bioinformatically. As shown in Figure S1E, all heterodimeric linker ligations, giving rise to chimeric PETs, are by definition nonspecific. Because random intermolecular associations in the test tube are expected to be comparable for linkers A and B, the frequency of random homo and heterodimeric linker ligations should also be equivalent. In our SMC1 ChIA-PET library, only 7% of pair-end ligations involved heterodimeric linkers (Table S1). Thus, we estimate that less than 14% of total homodimeric ligations are nonspecific. Following proximity ligation, samples were treated with Proteinase K and DNA was purified. An EcoP15I (NEB) digestion was performed at 37 degrees Celsius for 17 hours to

linearize the ligated chromatin fragments. The chromatin fragments were then immobilized on Dynabeads M280 Streptavidin beads. An End-Repair reaction was performed (Epicentre #ER81050), then As were added to the ends with Klenow treatment by rotating at 37 degrees Celsius for 35 minutes. Next, Illumina paired-end sequencing adapters were ligated on the ends and 18 cycles of PCR was performed. The Paired-End-Tag (PET) constructs were extracted from the ligation products and the PET templates were subjected to 50x50 paired-end sequencing using Illumina HiSeq 2000.

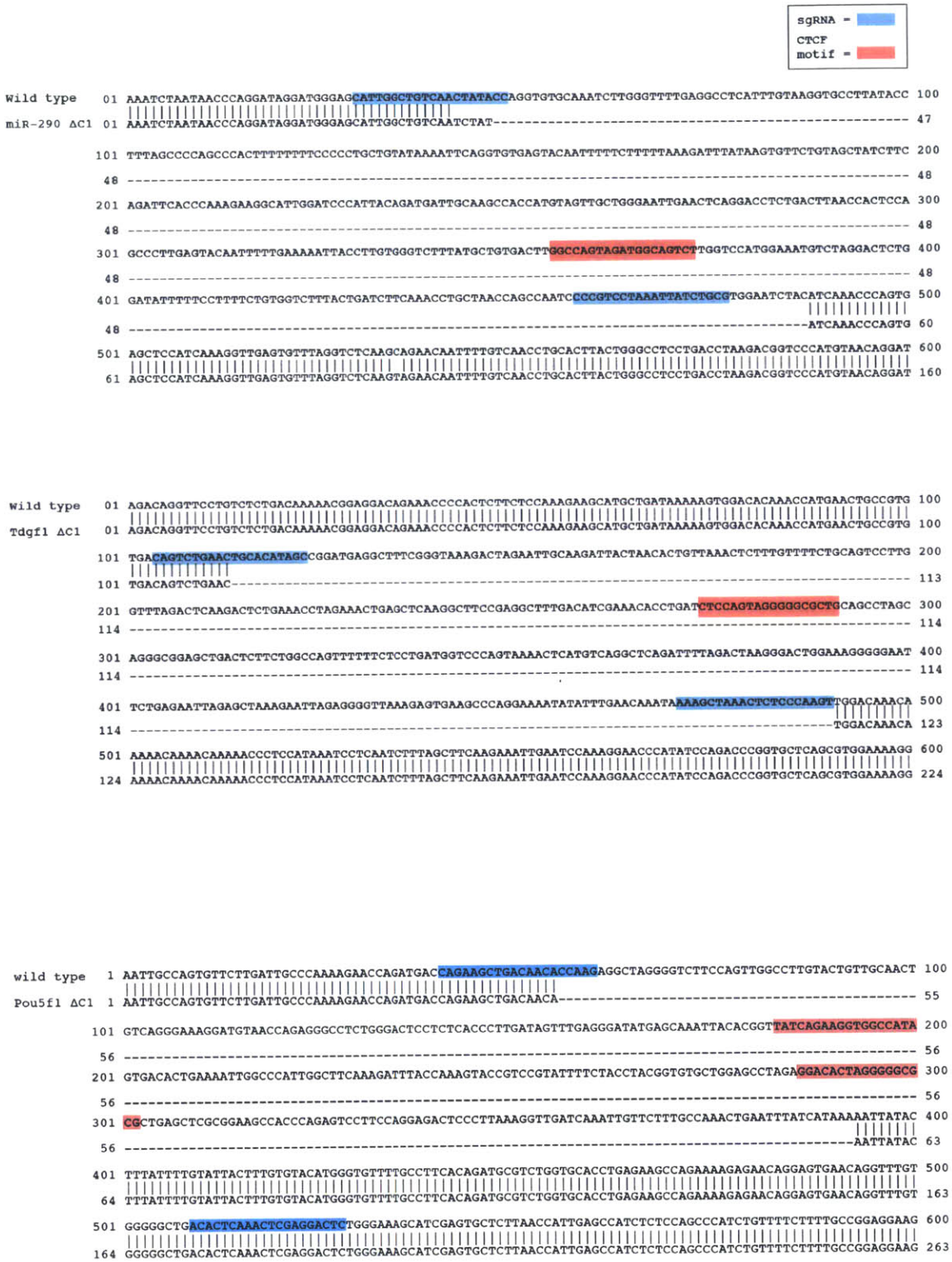
Genome Editing

The CRISPR/Cas9 system was used to create ESC lines with CTCF site deletions. Target-specific oligonucleotides were cloned into a plasmid carrying a codon-optimized version of Cas9 (pX330, Addgene: 42230). The genomic sequences complementary to guide RNAs in the genome editing experiments are:

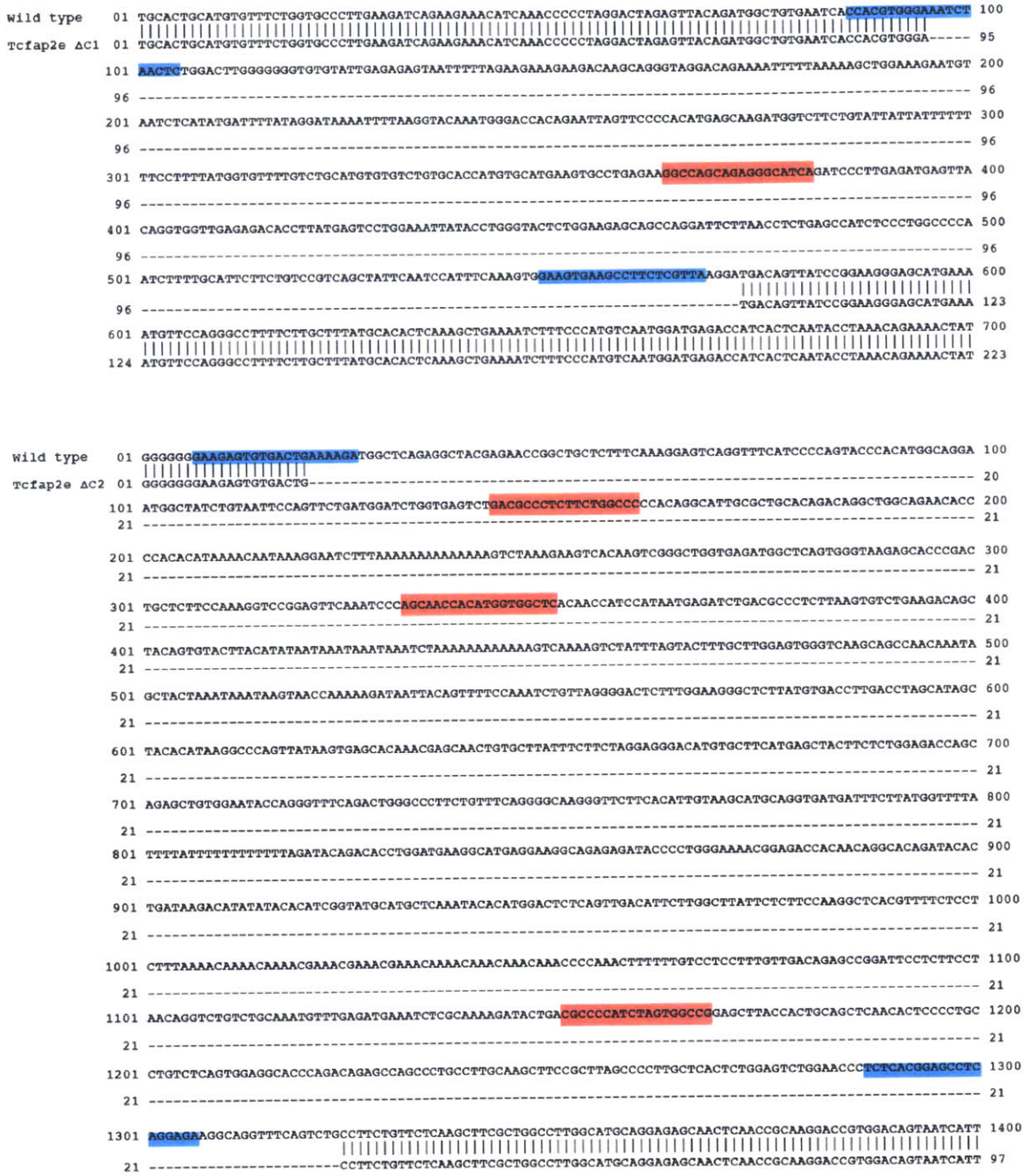
Name	Sequence
PRDM14_C1_up	ATGACATAATGAGATTCACG
PRDM14_C1_down	ACTGAAGTGGAAGGTGAGTG
PRDM14_C2_down	CGACCCACCTCCTAACCTTA
MIR290_C1_up	CATTGGCTGTCAACTATACC
MIR290_C1_down	CCCGTCCTAAATTATCTGCG
POU5F1_C1_up	CAGAAGCTGACAACACCAAG
POU5F1_C1_down	ACACTCAAACTCGAGGACTC
NANOG_C1_up	TTAAACACATCATAAGATGA
NANOG_C1_down	TGAACTACGTAGCAAGTTCC

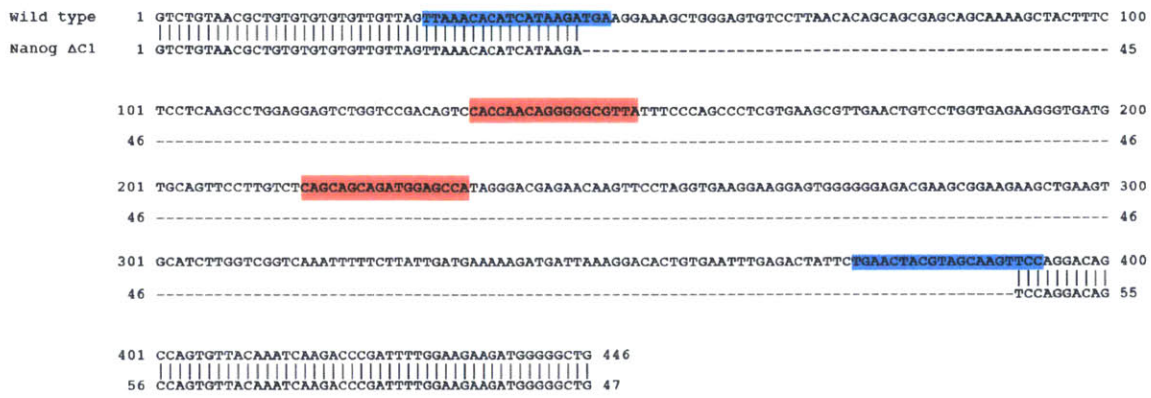
TDGF1_C1_up	CAGTCTGAACTGCACATAGC
TDGF1_C1_down	AAAGCTAAACTCTCCCAAGT
TCFAP2E_C1_up	CCACGTGGGAAATCTAACTC
TCFAP2E_C1_down	GAAGTGAAGCCTTCTCGTTA
TCFAP2E_C2_up	GAAGAGTGTGACTGAAAAGA
TCFAP2E_C2_down	TCTCACGGAGCCTCAGGAGA

Cells were transfected with two plasmids expressing Cas9 and sgRNA targeting regions around 200 basepairs up- and down-stream of the CTCF binding site, respectively. A plasmid expressing PGK-puroR was also co-transfected. Transfection was carried out with the X-fect reagent (Clontech) according to the manufacturer's instructions. One day after transfection, cells were re-plated on DR4 MEF feeder layers. One day after re-plating puromycin (2μg/ml) was added for three days. Subsequently, puromycin was withdrawn for three to four days. Individual colonies were picked, and genotyped by PCR. For the *Prdm14* (C1-2), *mir-290-295*, *Pou5f1* and *Nanog* SDs and *Tcfap2e* (C1) PD boundary CTCF site deletions, at least two independent clones were expanded and analyzed. Data on Figure 4, 5 and S4 were obtained from the analysis of a single representative clone for each genotype. The sequences of the deletion alleles in the used cell lines are listed below. The sites complementary to the sgRNAs are highlighted in a blue box, the CTCF motifs (JASPAR ID: M0A139.1) are highlighted in a red box.


sgRNA =
CTCF motif =
Wild type 01 AAATCTAATAACCCAGGATAGGATGGGAGCATTGGCTGTCAACTATACCAGGTGTGCAAATCTTGGGTTTTGAGGCCTCATTTGTAAGGTGCCTTATACC 100
miR-290 ΔC1 01 AAATCTAATAACCCAGGATAGGATGGGAGCATTGGCTGTCAATCTAT 47
101 TTTAGCCCCAGCCCACTTTTTTTTCCCCCTGCTGTATAAAATTCAGGTGTGAGTACAATTTTTCTTTTTAAAGATTTATAAGTGTTCTGTAGCTATCTTC 200
48 48
201 AGATTCACCCAAAGAAGGCATTGGATCCCATTACAGATGATTGCAAGCCACCATGTAGTTGCTGGGAATTGAACTCAGGACCTCTGACTTAACCACTCCA 300
48 48
301 GCCCTTGAGTACAATTTTTGAAAAATTACCTTGTGGGTCTTTATGCTGTGACTTGGCCAGTAGATGGCAGTCTTGGTCCATGGAAATGTCTAGGACTCTG 400
48 48
401 GATATTTTTCCTTTTCTGTGGTCTTTACTGATCTTCAAACCTGCTAACCAGCCAATCCCCGTCCTAAATTATCTGCGTGGAATCTACATCAAACCCAGTG 500
48 ATCAAACCCAGTG 60
501 AGCTCCATCAAAGGTTGAGTGTTTAGGTCTCAAGCAGAACAATTTTGTCAACCTGCACTTACTGGGCCTCCTGACCTAAGACGGTCCCATGTAACAGGAT 600
61 AGCTCCATCAAAGGTTGAGTGTTTAGGTCTCAAGTAGAACAATTTTGTCAACCTGCACTTACTGGGCCTCCTGACCTAAGACGGTCCCATGTAACAGGAT 160
Wild type 01 AGACAGGTTCCTGTCTCTGACAAAAACGGAGGACAGAAACCCCACTCTTCTCCAAAGAAGCATGCTGATAAAAAGTGGACACAAACCATGAACTGCCGTG 100
Tdgf1 ΔC1 01 AGACAGGTTCCTGTCTCTGACAAAAACGGAGGACAGAAACCCCACTCTTCTCCAAAGAAGCATGCTGATAAAAAGTGGACACAAACCATGAACTGCCGTG 100
101 TGACAGTCTGAACTGCACATAGCCGGATGAGGCTTTCGGGTAAAGACTAGAATTGCAAGATTACTAACACTGTTAAACTCTTTGTTTTCTGCAGTCCTTG 200
101 TGACAGTCTGAAC 113
201 GTTTAGACTCAAGACTCTGAAACCTAGAAACTGAGCTCAAGGCTTCCGAGGCTTTGACATCGAAACACCTGATCTCCAGTAGGGGGCGCTGCAGCCTAGC 300
114 114
301 AGGGCGGAGCTGACTCTTCTGGCCAGTTTTTTCTCCTGATGGTCCCAGTAAAACTCATGTCAGGCTCAGATTTTAGACTAAGGGACTGGAAAGGGGGAAT 400
114 114
401 TCTGAGAATTAGAGCTAAAGAATTAGAGGGGTTAAAGAGTGAAGCCCAGGAAAATATATTTGAACAAATAAAAGCTAAACTCTCCCAAGTTGGACAAACA 500
114 TGGACAAACA 123
501 AAAACAAAACAAAAACCCTCCATAAATCCTCAATCTTTAGCTTCAAGAAATTGAATCCAAAGGAACCCATATCCAGACCCGGTGCTCAGCGTGGAAAAGG 600
124 AAAACAAAACAAAAACCCTCCATAAATCCTCAATCTTTAGCTTCAAGAAATTGAATCCAAAGGAACCCATATCCAGACCCGGTGCTCAGCGTGGAAAAGG 224
wild type 1 AATTGCCAGTGTTCTTGATTGCCCAAAAGAACCAGATGACCAGAAGCTGACAACACCAAGAGGCTAGGGGTCTTCCAGTTGGCCTTGTACTGTTGCAACT 100
Pou5f1 ΔC1 1 AATTGCCAGTGTTCTTGATTGCCCAAAAGAACCAGATGACCAGAAGCTGACAACA 55
101 GTCAGGGAAAGGATGTAACCAGAGGGCCTCTGGGACTCCTCTCACCCTTGATAGTTTGAGGGATATGAGCAAATTACACGGTTATCAGAAGGTGGCCATA 200
56 56
201 GTGACACTGAAAATTGGCCCATTGGCTTCAAAGATTTACCAAAGTACCGTCCGTATTTTCTACCTACGGTGTGCTGGAGCCTAGAGGACACTAGGGGGCG 300
56 56
301 CGCTGAGCTCGCGGAAGCCACCCAGAGTCCTTCCAGGAGACTCCCTTAAAGGTTGATCAAATTGTTCTTTGCCAAACTGAATTTATCATAAAAATTATAC 400
56 AATTATAC 63
401 TTTATTTTGTATTACTTTGTGTACATGGGTGTTTTGCCTTCACAGATGCGTCTGGTGCACCTGAGAAGCCAGAAAAGAGAACAGGAGTGAACAGGTTTGT 500
64 TTTATTTTGTATTACTTTGTGTACATGGGTGTTTTGCCTTCACAGATGCGTCTGGTGCACCTGAGAAGCCAGAAAAGAGAACAGGAGTGAACAGGTTTGT 163
501 GGGGGCTGACACTCAAACTCGAGGACTCTGGGAAAGCATCGAGTGCTCTTAACCATTGAGCCATCTCTCCAGCCCATCTGTTTTCTTTTGCCGGAGGAAG 600
164 GGGGGCTGACACTCAAACTCGAGGACTCTGGGAAAGCATCGAGTGCTCTTAACCATTGAGCCATCTCTCCAGCCCATCTGTTTTCTTTTGCCGGAGGAAG 263


Wild type
Prdm14 ΔC1-2


Wild type 01 TGCACTGCATGTGTTTCTGGTGCCCTTGAAGATCAGAAGAAACATCAAACCCCCTAGGACTAGAGTTACAGATGGCTGTGAATCACCACGTGGGAAATCT 100
Tcfap2e ΔC1 01 TGCACTGCATGTGTTTCTGGTGCCCTTGAAGATCAGAAGAAACATCAAACCCCCTAGGACTAGAGTTACAGATGGCTGTGAATCACCACGTGGGA----- 95
101 AACTCTGGACTTGGGGGGGTGTGTATTGAGAGAGTAATTTTTAGAAGAAAGAAGACAAGCAGGGTAGGACAGAAAATTTTTAAAAAGCTGGAAAGAATGT 200
96 ---------- 96
201 AATCTCATATGATTTTATAGGATAAAATTTTAAGGTACAAATGGGACCACAGAATTAGTTCCCCACATGAGCAAGATGGTCTTCTGTATTATTATTTTTT 300
96 ---------- 96
301 TTCCTTTTATGGTGTTTTGTCTGCATGTGTGTCTGTGCACCATGTGCATGAAGTGCCTGAGAAGGCCAGCAGAGGGCATCAGATCCCTTGAGATGAGTTA 400
96 ---------- 96
401 CAGGTGGTTGAGAGACACCTTATGAGTCCTGGAAATTATACCTGGGTACTCTGGAAGAGCAGCCAGGATTCTTAACCTCTGAGCCATCTCCCTGGCCCCA 500
96 ---------- 96
501 ATCTTTTGCATTCTTCTGTCCGTCAGCTATTCAATCCATTTCAAAGTGGAAGTGAAGCCTTCTCGTTAAGGATGACAGTTATCCGGAAGGGAGCATGAAA 600
96 ----------TGACAGTTATCCGGAAGGGAGCATGAAA 123
601 ATGTTCCAGGGCCTTTTCTTGCTTTATGCACACTCAAAGCTGAAAATCTTTCCCATGTCAATGGATGAGACCATCACTCAATACCTAAACAGAAAACTAT 700
124 ATGTTCCAGGGCCTTTTCTTGCTTTATGCACACTCAAAGCTGAAAATCTTTCCCATGTCAATGGATGAGACCATCACTCAATACCTAAACAGAAAACTAT 223
Wild type 01 GGGGGGGAAGAGTGTGACTGAAAAGATGGCTCAGAGGCTACGAGAACCGGCTGCTCTTTCAAAGGAGTCAGGTTTCATCCCCAGTACCCACATGGCAGGA 100
Tcfap2e ΔC2 01 GGGGGGGAAGAGTGTGACTG---------- 20
101 ATGGCTATCTGTAATTCCAGTTCTGATGGATCTGGTGAGTCTGACGCCCTCTTCTGGCCCCCACAGGCATTGCGCTGCACAGACAGGCTGGCAGAACACC 200
21 ---------- 21
201 CCACACATAAAACAATAAAGGAATCTTTAAAAAAAAAAAAAAAGTCTAAAGAAGTCACAAGTCGGGCTGGTGAGATGGCTCAGTGGGTAAGAGCACCCGAC 300
21 ---------- 21
301 TGCTCTTCCAAAGGTCCGGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCATCCATAATGAGATCTGACGCCCTCTTAAGTGTCTGAAGACAGC 400
21 ---------- 21
401 TACAGTGTACTTACATATAATAAATAAATAAATCTAAAAAAAAAAAAAGTCAAAAGTCTATTTAGTACTTTGCTTGGAGTGGGTCAAGCAGCCAACAAATA 500
21 ---------- 21
501 GCTACTAAATAAATAAGTAACCAAAAAGATAATTACAGTTTTCCAAATCTGTTAGGGGACTCTTTGGAAGGGCTCTTATGTGACCTTGACCTAGCATAGC 600
21 ---------- 21
601 TACACATAAGGCCCAGTTATAAGTGAGCACAAACGAGCAACTGTGCTTATTTCTTCTAGGAGGGACATGTGCTTCATGAGCTACTTCTCTGGAGACCAGC 700
21 ---------- 21
701 AGAGCTGTGGAATACCAGGGTTTCAGACTGGGCCCTTCTGTTTCAGGGGCAAGGGTTCTTCACATTGTAAGCATGCAGGTGATGATTTCTTATGGTTTTA 800
21 ---------- 21
801 TTTTATTTTTTTTTTTTAGATACAGACACCTGGATGAAGGCATGAGGAAGGCAGAGAGATACCCCTGGGAAAACGGAGACCACAACAGGCACAGATACAC 900
21 ---------- 21
901 TGATAAGACATATATACACATCGGTATGCATGCTCAAATACACATGGACTCTCAGTTGACATTCTTGGCTTATTCTCTTCCAAGGCTCACGTTTTCTCCT 1000
21 ---------- 21
1001 CTTTAAAACAAAACAAAACGAAACGAAACGAAACAAAACAAACAAACCCCAAACTTTTTTGTCCTCCTTTGTTGACAGAGCCGGATTCCTCTTCCT 1100
21 ---------- 21
1101 AACAGGTCTGTCTGCAAATGTTTGAGATGAAATCTCGCAAAAGATACTGACGCCCCATCTAGTGGCCGGAGCTTACCACTGCAGCTCAACACTCCCCTGC 1200
21 ---------- 21
1201 CTGTCTCAGTGGAGGCACCCAGACAGAGCCAGCCCTGCCTTGCAAGCTTCCGCTTAGCCCCTTGCTCACTCTGGAGTCTGGAACCCTCTCACGGAGCCTC 1300
21 ---------- 21
1301 AGGAGAAGGCAGGTTTCAGTCTGCCTTCTGTTCTCAAGCTTCGCTGGCCTTGGCATGCAGGAGAGCAACTCAACCGCAAGGACCGTGGACAGTAATCATT 1400
21 ----------CCTTCTGTTCTCAAGCTTCGCTGGCCTTGGCATGCAGGAGAGCAACTCAACCGCAAGGACCGTGGACAGTAATCATT 97



Gene Expression Analysis

ESC lines were split off MEFs for two passages. RNA was isolated using Trizol reagent (Invitrogen) or RNeasy purification kit (Promega), and reverse transcribed using oligo-dT primers and SuperScript III reverse transcriptase (Invitrogen) according to the manufacturers´ instructions. Quantitative real-time PCR was performed on a 7000 AB Detection System using the following Taqman probes, according to the manufacturer´s instructions (Applied Biosystems):

Gapdh: Mm99999915_g1

Prdm14: Mm01237814_m1

Slco5a1: Mm00556042_m1

Pou5f1: Mm00658129_gH

H2-Q10: Mm01275264_g1

Tcf19: Mm00508531_m1

Mmu-mir-292b: Mm03307733_pri

Nlrp12: Mm01329688_m1

Myadm: Mm01329822_m1

AU018091: Mm01329669_m1

Nanog: Mm02019550_s1

Dppa3: Mm01184198_g1

Tdgf1: Mm03024051_g1

Gm590: Mm01250263_m1

Lrrc2: Mm01250173_m1

Rtp3: Mm00462169_m1

Tcfap2e: Mm01179789_m1

Psmb2: Mm00449477_m1

Ncdn: Mm00449525_m1

Sox2: Mm03053810_s1

Pax6: Mm00443081_m1

Gata6: Mm00802636_m1

Sox17: Mm00488363_m1

Based on RNA-seq data (Shen et al. 2012), the genes are expressed at the following levels prior to deletion of the CTCF site:

Pou5f1: 79.4 RPKM (rank among 24,827 Refseq transcripts: 232, top 1%)

Prdm14: 2.21 RPKM (rank: 9,745, 39th%)

Slco5a1: 0.93 RPKM (rank: 12,277, 50th%)

miR-295: 18.9 RPKM (rank: 1,902, 8th%)

H2-Q10: 0.48 RPKM (rank: 13,782, 56th%)

Tcf19: 1.03 RPKM (rank: 12,011, 49th%)

Nlrp12: 0.06 RPKM (17,108, 69th%)

AU018091: 17.1 RPKM (rank: 2,150, 9th%)

Myadm: 14.6 RPKM (mean of multiple splice isoforms) (rank: 2610, 11th%)

Dppa3: 25 RPKM (rank: 1,320, 5th%)

Tdgf1: 92 RPKM (rank: 167, top 1%)

Lrrc2: 1.2 RPKM (rank: 10,292, 42nd%)

Rtp3: 0.01 RPKM (rank: 14,587 59th),

Sox2: 122 RPKM (rank: 100, top 1%)

Nanog: 122 RPKM (rank: 99, top 1%)

Pax6: 0.07 RPKM (rank: 16,941, 68th%)

Gata6: 0.25 RPKM (rank: 14,981, 60th%)

Sox17: 0.15 RPKM (rank: 15,754, 64th%)

Psmb2: 85 RPKM (rank: 203, top 1%)

Tcfap2e: 0.19 RPKM (rank: 15,402, 62nd%)

Ncdn: 3.19 RPKM (rank: 8,388, 24th%)

ChIP-Seq Illumina Sequencing and Library Generation

Purified DNA from a H3K27me3 ChIP was used to prepare a library for Illumina sequencing. The library was prepared following the Illumina TruSeq DNA Sample Preparation v2 kit protocol as previously described (Whyte et al. 2012).

3C assays

For each sample, 2X10^{7} ESCs cells were crosslinked with 1%formaldehyde for 20 min at RT. The reaction was quenched by the addition of 125mM glycine for 5 min at RT. Crosslinked ESCs were washed with PBS and resuspended in 10ml lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% NP40 and proteinase inhibitors) and lysed with a Dounce homogenizer. Following BglII digestion overnight, 3C-ligated DNA was prepared as previously described (Lieberman-

Aiden et al. 2009). The 3C interactions at the *miR-290-295* and *Pou5f1* loci (Figure S4A, S4B) were analyzed by quantitative real-time PCR using custom Taqman probes as previously described (Xu et al. 2011). The amount of DNA in the qPCR reactions was normalized across 3C libraries using a custom Taqman probe directed against the *Actb* locus. Primer sequences are listed below.

Target region	*Primer name*	*Sequence (5´-3´)*
Nlrp12 promoter	Nlrp12 R	CACATCTTCAAAGCAAACACTATTGTT
Nlrp12 Taqman probe	Nlrp12 Probe	TCTCCTACCCATTGCTTCTCTGCTACCTGC
SE region 1	Nlrp12 eF1	TTCCTGGAACCTGGGCAA
SE region 2	Nlrp12 eF2	TGATACAGCACAGCTTTCCTTCA
SE region 3	Nlrp12 eF3	CAGATTTTTTATTTCCTTCAGTTCTGTG
H2-Q10 promoter	H2Q10 F	AGGGCTCACCTTCAGTCAAGTT
SE region	H2Q10 R	AGGATGGCTCAGCGGTTAAG
H2-Q10 Taqman probe	H2Q10 probe	CGGCCTGTCTACTTTAGCCTCAGACTCCA
Actin	Actin-F	GGG AGT GACTCT CTG TCC ATT CA
Actin	Actin-R	ATT TGT GTG GCCTCT TGT TTG A
Actin Taqman probe	Actin probe	TCC AGG CCC CGC GTG TCC

F, and R denote forward and reverse primers, respectively.

Bioinformatics Analysis

ChIP-seq Data Analysis

All ChIP-Seq data sets were aligned using Bowtie (version 0.12.2) (Langmead et al. 2009) to build version MM9 of the mouse genome with parameter -k 1 -m 1 -n 2. Data sets used in this manuscript can be found in Table S16. We used the MACS version 1.4.2 (model-based analysis of ChIP-seq) (Zhang et al. 2008) peak finding algorithm to identify regions of ChIP-seq enrichment over input DNA control. A p value threshold of enrichment of 1e-09 was used for all data sets. For the histone modification H3K27me3 whose signal tends to be broad across large genomic regions, we used MACS (Zhang et al. 2008) with the parameter "-p 1e-09 -no-lambda –no-model". UCSC Genome Browser (Kent et al. 2002) tracks were generated using MACS wiggle outputs with parameters "-w –S –space=50".

SMC1 ChIP-seq Enrichment Heatmap

Figure 1B, S1A, and S1B shows the average ChIP-seq read density (r.p.m./bp) of different factors at the indicated sets of regions. The average ChIP-seq in 50 bp bin was calculated and drawn. In Figure 1B, +/- 5 kb from the center of the SMC1-enriched region was interrogated. In Figure S1A, the enriched regions of OSN, MED1, and MED12 were merged together if overlapping by 1 bp. For each of the merged regions, +/- 5 kb from the center of the merged region was interrogated. On Figure S1B, +/- 5 kb from the center of the CTCF enriched region was interrogated.

Gene Sets and Classification of Gene Transcriptional State in ESCs

All gene-centric analyses in ESCs were performed using mouse (mm9/NCBI37) RefSeq annotations downloaded from the UCSC genome browser (genome.ucsc.edu). For counting purposes and for assignment of enhancers to target genes, we collapsed multiple identical TSS into one gene-level TSS. Genes were separated into classes of activity as follows:

A gene was defined as active if an enriched region for either H3K4me3 or RNA Pol II was located within +/- 2.5 kb of the TSS and lacked an enriched region for H3K27me3 therein. H3K4me3 is a histone modification associated with transcription initiation (Guenther et al. 2007).

A gene was defined as Polycomb-occupied if an enriched region for H3K27me3 (representing Polycomb complexes) but not RNA Pol II was located within +/- 2.5 kb of the TSS. H3K27me3 is a histone modification associated with Polycomb complexes (Boyer et al. 2006, Lee et al. 2006).

A gene was defined as silent if H3K4me3, H3K27me3, or RNA Pol II enriched regions was absent from +/- 2.5 kb of the TSS.

Remaining genes to which we were unable to assign a state were left as unclassified. Overall, there were 15,312 unique active TSSs, 1,091 unique Polycomb-occupied TSSs, 8,477 unique silent TSSs, and 616 unclassified TSSs in mouse ES cells.

Defining Active Enhancers in ESCs

Co-occupancy of ESC genomic sites by the OCT4, SOX2, and NANOG transcription factors is highly predictive of enhancer activity (Chen et al. 2008) and Mediator is typically associated with these sites (Kagey et al. 2010). We first pooled the reads of ChIP-seq profiles of transcription factors OCT4, SOX2, and NANOG, which were performed in parallel, to create a merged "OSN" ChIP-seq experiment (Whyte et al. 2013). These reads were processed by MACS to create an OSN binding profile for visualization. To define active enhancers, we first identified enriched regions for the merged "OSN" ChIP-seq read pool, and for both Mediator complex components MED1 and MED12 using MACS. Then we used the union of these five sets of enriched ChIP-Seq regions that fell outside of promoters (e.g., a region not overlapping with ± 2.5 kb region flanking the RefSeq transcriptional start sites) as putative enhancers.

SMC1 ChIA-PET Processing

All ChIA-PET datasets were processed with a method adapted from a previous computational pipeline (Li et al. 2010). The raw sequences were analyzed for linker barcode composition and separated into non-chimeric PET sequences with homodimeric linkers (AA or BB linkers) derived from specific ligation products, or chimeric PET sequences (AB linkers) with heterodimeric linker derived from nonspecific ligation products. We trimmed the 3' end of PET sequences after a perfect match of the first 10nt of the linker sequences (Linker A with CTGCTGTCCG; Linker B with CTGCTGTCAT). After removing the linkers, only the 5' ends of the trimmed PET sequences of at least 27bp were retained,

because the restriction enzyme EcoP151 cuts 27bp away from its recognition sequence. The sequences of the two ends of PETs were separately mapped to the mm9 mouse genome using the bowtie algorithm with the option "-k 1 -m 1 -v 1" (Langmead et al. 2009). These criteria retained only the uniquely mapped reads, with at most a single mismatch for further analysis. Aligned reads were paired with mates using read identifiers and, to remove PCR bias artifacts, were filtered for redundancy: PETs with identical genomic coordinates and strand information at both ends were collapsed into a single PET. The PETs were further categorized into intrachromosomal PETs, where the two ends of a PET were on the same chromosome, and interchromosomal PETs, where the two ends were on different chromosomes. The end read positions of all non-chimeric PETs were used to call PET peaks that represent local enrichment of the PET sequence coverage by using MACS 1.4.2 (Zhang et al. 2008) with the parameters "-p 1e-09 -no-lambda –no-model".

Chimeric Versus Non-chimeric PET Quality Assessment

Chimeric PETs with heterodimeric linkers can be used to estimate the degree of noise in the ChIA-PET dataset. Since only 7% of paired-end ligations involved heterodimeric linkers (Table S1), we estimated that less than 14% of total homodimeric ligations were nonspecific. We also counted the chimeric PET sequences that overlapped with PET peaks at both ends by at least 1bp. These chimeric PET sequences represented "non-specific" chromatin interactions. We found that more than 99.8% "non-specific" chromatin interactions derived from

chimeric PET sequences overlapping with PET peaks had only 1 chimeric PET; 0.1% "non-specific" interactions had 2 chimeric PETs. We thus used a 3 PET cut-off for our high-confidence interactions (Figure S1F). Since contact frequency is expected to inversely scale with genomic distance, we examined the relationship between PET frequencies over genomic distance between the two ends of intra-chromosomal PET sequences. The frequency of non-chimeric PETs with homodimeric linkers was plotted over genomic span in increments of 100bp (Figure S1E). The scatter plot suggested two populations within intra-chromosomal PETs and showed that the vast majority of these PETs were within 4 kb (Figure S1E). We thus used a 4 kb cutoff to remove those PET sequences that may originate from self-ligation of DNA ends from a single chromatin fragment in the ChIA-PET procedure. In contrast, chimeric PETs with heterodimeric linkers did not show an inverse relationship with genomic distance (Figure S1E, Table S1).

Creation of High-Confidence ChIA-PET Interactions

To identify long-range chromatin interactions, we first removed intra-chromosomal PETs of length < 4 kb because these PETs may originate from self-ligation of DNA ends from a single chromatin fragment in the ChIA-PET procedure (Figure S1E, see above). We next identified PETs that overlapped with PET peaks at both ends by at least 1bp. Operationally, these PETs were defined as putative interactions. Applying a statistical model based upon the hypergeometric distribution identified high-confidence interactions, representing

high-confidence physical linking between the PET peaks. Specifically, the numbers of PET sequences that overlapped with PET peaks at both ends as well as the number of PETs within PET peaks at each end were counted. The PET count between two PET peaks represented the frequency of the chromatin interaction between the two genomic locations. A hypergeometric distribution was used to determine the probability of seeing at least the observed number of PETs linking the two PET peaks. A background distribution of interaction frequencies was then obtained through the random shuffling of the links between two ends of PETs, and a cutoff threshold for calling significant interactions was set to the corresponding p-value of the most significant proportion of shuffled interactions (at an FDR of 0.01). This method yielded similar number of interactions as the correction of p-values by the Benjamini-Hochberg procedure (Benjamini 1995) to control for multiple hypothesis testing. Operationally, the pairs of interacting sites with three independent PETs were defined as high-confidence interactions in the SMC1 ChIA-PET merged dataset (Table S15), and with two independent PETs in the individual SMC1 ChIA-PET replicates (Table S13, S14).

Saturation Analysis of ChIA-PET Library

To determine the degree of saturation within our ChIA-PET library (Figure S1H), we modeled the number of sampled genomic positions as a function of sequencing depth by the Michaelis-Menten model. Intrachromosomal PETs with a distance span above our self-ligation cutoff of 4 kb were subsampled at varying

depths, and the number of unique genomic positions (defined as the start and end coordinates of the paired PETs) that they occupy were counted. Model fitting using non-linear least-squares regression suggested that we have sampled approximately 70 % of the available intrachromosomal PET space, encompassing 2.22 /3.17 million positions (Figure S1H).

We considered whether ChIA-PET data limitations might limit detection of longer-range interactions. If sparseness of data were a significant problem, resulting in under-calling of long-range interactions, we would likely miss previously detected long-range interactions. Instead, we detect previously known long-range interactions, e.g. the interaction between Sonic Hedgehog (Shh) and its enhancer in the intron of the nearby Lmbr1 gene (1 Mb away), interactions between the HoxD gene cluster and its distal regulatory sequences (>300 kb away), and interactions between the HoxA gene cluster and its distal regulatory sequences (>500 kb away) (Lehoczky, Williams, and Innis 2004, Lettice et al. 2003, Spitz, Gonzalez, and Duboule 2003).

Assignment of Interactions to Regulatory Elements

To identify the association of long-range chromatin interactions to different regulatory elements, we assigned the PET peaks of interactions to different regulatory elements, including active enhancers, promoters (+/- 2.5 kb of the Refseq TSS), and CTCF sites. Operationally, an interaction was defined as associated with the regulatory element if one of the two PET peak of the interaction overlapped with the regulatory element by at least 1 base-pair.

Assignment of Enhancers to Genes

Our analysis identified 2,921 high-confidence interactions involving an enhancer (contains an OCT4/SOX2/NANOG/MED1/MED12 enriched region and is not located within +/-2.5 kb of an annotated TSS) and a promoter (+/- 2.5 kb of an annotated TSS) (Figure S1C, Table S15). Each high-confidence interaction, as defined above, is required to be connected by three PET peaks. A large majority (81%) of these enhancer-promoter interactions (2071/2921 interactions) involved an active gene (H3K4me3 or RNA Pol II but not H3K27me3 enriched regions), while 302 interactions involved a Polycomb-occupied gene (H3K27me3) and 229 interactions involved a silent gene (absence of H3K4me3, RNA Pol II and H3K27me3 enriched regions). We identified 216 enhancer-promoter interactions that involved super-enhancers (Table S4), as defined in (Hnisz et al. 2013, Whyte et al. 2013)

The high-confidence enhancer-promoter interactions were used to assign super-enhancers and typical enhancers to their target genes (Table S4, S5). Multiple enhancer constituents that are in close proximity can be computationally stitched together into enhancer regions (true for typical and super-enhancers) as described previously (Hnisz et al. 2013, Whyte et al. 2013). We identified high-confidence interactions overlapping with a super-enhancer or typical enhancer region at one end and a TSS (+/- 2.5 kb of a TSS) at the other end (Table S4, S5). For super-enhancers with sufficient interaction data, we found that 83% of enhancer assignments to the nearest active gene (including Polycomb-occupied

genes) were confirmed/supported by high-confidence interactions. For typical enhancers with sufficient interaction data, we found that 87% of enhancer assignments to the nearest active gene (including Polycomb-occupied genes) were confirmed/supported by high-confidence interaction data.

Heatmap Representation of High-Confidence ChIA-PET Interactions at Topologically Associating Domains (TADs)

Genome-wide average representations of ChIA-PET interactions were created by mapping high-confidence ChIA-PET interactions across TADs (Dixon et al. 2012) (Figure 2D). All ~2,200 TADs plus their upstream and downstream flanking regions (10% of the size of the domain) were aligned and each split into 60 equally-sized bins. To calculate interaction density in each TAD, we first filtered high-confidence interactions by requiring they were completely contained within the genomic region of the TAD and its flanking regions defined above. We next counted the interaction frequency between any two bins in each TAD to produce a 60 by 60 interaction matrix using a method as previously described in Dixon et al., 2012 The numbers in the interaction matrices represent interaction frequencies at the diagonals originating from two bins on the x- and y- axis. Average interaction frequencies across ~2,200 TAD interaction matrices were calculated. The upper triangular matrix of the average interaction frequencies was displayed in the units of interactions per bin in Figure 2D.

Definition of Super-enhancer Domains and Polycomb-repressed Domains

Typical enhancer and super-enhancer regions in murine embryonic stem cells were described previously (Hnisz et al. 2013, Whyte et al. 2013), and their genomic coordinates were downloaded (Table S4, S5). The 231 super-enhancers were assigned to genes with a combination of ChIA-PET interactions and proximity to their nearest active transcriptional start sites (TSSs). We first used high-confidence SMC1 PET interactions (FDR 0.01, 3 PETs) between super-enhancers and TSS regions (+/- 2.5 kb of a TSS) to identify their target genes. When super-enhancers did not have PET interactions to any TSS regions, they were assigned the nearest active TSSs (including Polycomb-occupied genes) by proximity. Super-enhancers and the TSS regions (+/- 2.5 kb of a TSS) of their target genes are considered as SE-gene units. All 231 super-enhancers were assigned to target genes with this method. This approach resulted in a total of 302 SE-gene units because a SE occasionally interacted with multiple genes. We next identified SMC1 PET interactions between two CTCF-enriched regions (regardless of whether these CTCF regions were at promoters or enhancers) that encompass these SE-gene units, which we called super-enhancer domains—we call these regions "CTCF-CTCF PET interactions." The CTCF-CTCF PET interactions defining super-enhancer domains were required to encompass the TSS regions (+/- 2.5 kb of a TSS) and the super-enhancer for each SE-gene unit. When multiple nested CTCF-CTCF PET interactions encompassed a SE-gene unit, we used the smallest CTCF-CTCF PET interactions for simplicity. We identified 193 Super-enhancer Domains (SDs) containing a total of 191 super-enhancers. We noted that the boundaries of super-enhancer are sensitive to the

algorithm that computationally defines super-enhancers. For 4 super-enhancers, one super-enhancer constituent out of multiple constituent enhancers that define the super enhancers fall outside of the CTCF-CTCF PET interactions. These 4 CTCF PET interactions encompass the target gene TSS regions (+/- 2.5 kb of a TSS) and more than 50% of the genomic space covered by the super-enhancer. Therefore, we qualified these 4 CTCF-CTCF PET interactions as Super-enhancer Domains. Thus, we identified a total of 197 Super-enhancer Domains (SDs) containing a total of 197 boundary CTCF-CTCF PET interactions and 195 super-enhancers (Table S7, S8). For the ~15% super-enhancers that did not qualify for occurrence within a SD by using the high confidence ChIA-PET data, the interaction dataset (not the high confidence data) shows that all but one of these super-enhancers are located within CTCF-CTCF loops co-bound by cohesin.

We also performed the same computational analyses for the 8,563 typical enhancers. We found that only 48% (4128/8563) typical-enhancers are contained in CTCF-CTCF topological structures similar to SDs.

Developmental regulators in embryonic stem cells frequently exhibit extended binding of Polycomb complex at their promoters spanning 2-35 kb from their promoters (Lee et al. 2006, Boyer et al. 2006). We thus focused on those Polycomb-occupied TSSs that showed enrichment of H3K27me3 spanning greater than 2 kb in size. This distance cutoff was based on analyses performed in (Lee et al. 2006). We noted that ~60% H3K27me3 regions called by MACS had neighboring H3K27me3 regions within 2 kb. In order to accurately capture

the large genomic regions that show enrichment of H3K27me3 signal, we first merged the H3K27me3 regions that were within 2 kb of each other. 546 genes, including 203 encoding transcription factors, showed enrichment of H3K27me3 spanning greater than 2 kb at their promoters. We next identified high confidence CTCF-CTCF PET interactions that encompassed the H3K27me3 regions of these 546 genes at promoters. When multiple nested CTCF-CTCF PET interactions encompassed the H3K27me3 regions, we took the smallest CTCF-CTCF PET interactions for simplicity. We identified 349 Polycomb Domains (PDs) containing a total of 349 boundary CTCF-CTCF PET interactions and 380 Polycomb-associated genes (Table S10, S11).

Support for SD and PD Structures from Published Datasets

The existence of Super-enhancer Domains and Polycomb-repressed Domains was supported by evidence from published CTCF ChIA-PET datasets (GSE28247) (Handoko et al. 2011). We applied our ChIA-PET processing method to the published CTCF ChIA-PET dataset to identify unique PETs. We then counted the instances where a high-confidence CTCF-CTCF boundary interaction from our ChIA-PET dataset showed a minimum 80% reciprocal overlap with the span of a unique PET from the CTCF ChIA-PET dataset, e.g. 80% of a high-confidence SD boundary interaction region is in common with a CTCF ChIA-PET unique PET and vice versa. To accomplish this, we used BEDtools intersect with parameters –f 0.8 –r –u. We found that 34% (6770/20080) of our CTCF-CTCF interactions were confirmed by a unique PET

within the CTCF ChIA-PET dataset, 33% (65/197) of our SD boundary interactions were confirmed by a unique PET within the CTCF ChIA-PET dataset, and 33% (115/349) of our PD boundary interactions were confirmed by a unique PET within the CTCF ChIA-PET dataset (Table S6).

Most Super-enhancer Domains and Polycomb-repressed Domains are distinct from the previously described Topologically Associating Domains (TADs). We compared Super-enhancer Domains and Polycomb-repressed Domains to TADs by counting the instances where a Super-enhancer Domain or a Polycomb-repressed Domain showed a minimum 80% reciprocal overlap with a TAD. 3% (5/197) of our SDs and 4% (13/349) of our PD have an 80% reciprocal overlap with a TAD (Dixon et al. 2012). 8% (16/197) of our SDs and 9% (30/349) of our PD have an 80% reciprocal overlap with a TAD (Filippova et al., 2014) (Table S6).

The existence of enhancer-promoter and enhancer-enhancer interactions was supported by evidence from published RNA PolII ChIA-PET datasets (Kieffer-Kwon et al. 2013). We applied our ChIA-PET processing method to the published Pol2 ChIA-PET dataset to identify unique PETs. We then counted the instances where a high-confidence enhancer-promoter or enhancer-enhancer interaction from our Smc1 ChIA-PET dataset showed a minimum 80% reciprocal overlap with a unique PET from the Pol2 ChIA-PET dataset, e.g. 80% of an enhancer-promoter interaction region is in common with a Pol2 ChIA-PET unique PET and vice versa. We found that 82% (2,402/2,921) of our enhancer-promoter interactions were confirmed by a unique PET within the Pol2 ChIA-PET dataset,

and 73% (1,969/2,700) of our enhancer-enhancer interactions were confirmed by a unique PET within the Pol2 ChIA-PET dataset (Table S6).

Several types of structural domains have been previously described, and we expect our interactions to occur largely within their boundaries. Thus, we determined how many of our interactions spanned a boundary. Topologically Associating Domains (TADs) (Dixon et al. 2012) were determined using Hi-C in mouse ESCs; 6% (1,354/23,739) of high-confidence intrachromosomal cohesin-mediated interactions cross a TAD boundary. LOCK (large organized chromatin K9 modification) domains were determined using ChIP data (Wen et al. 2009); 4% (1,053/23,739) of high-confidence, intrachromosomal cohesin-mediated interactions cross a LOCK boundary. Lamin-associated domains (LADs) were determined using DamID (Meuleman et al. 2013); 5% (1,180/23,739) of high-confidence intrachromosomal cohesin-mediated interactions cross a LAD boundary (Table S6).

Meta Representations of ChIP-Seq Occupancy at Super-Enhancer Domains and Polycomb Domains

Genome-wide average "meta" representations of ChIP-seq occupancy of different factors were created by mapping ChIP-seq read density to different sets of regions (Figure 3C, Figure 5C). All regions within each set were aligned and the average ChIP-Seq factor density in each bin was calculated to create a meta genome-wide average in units of rpm/bp. For super-enhancers, each super-enhancer or their corresponding flanking region (+/- 3 kb) was split into 100

equally-sized bins. This split all super-enhancer regions, regardless of their size, into 300 bins. For the target genes within SDs or PDs, we created three regions: upstream, gene body and downstream. 80 equally-sized bins divided the -2000 to 0 promoter region, 200 equally-sized bins divided the length of the gene body, and 80 equally-sized bins divided the 0 to + 2 kb downstream region. For SMC1 and CTCF sites at the SD, PD, and TAD borders, flanking regions (+/- 2 kb) around the center of CTCF sites were aligned and split into 40 equally-sized bins.

Heatmap representations of ChIP-seq occupancy of different factors were created by mapping ChIP-seq read density to the super-enhancer and their target genes in Super-enhancer Domains (Figure 3E). We created three types of regions: SD and their corresponding flanking regions(+/- 10 kb). We divided the upstream and downstream flanking regions into 10 equally-sized bins each. We divided the SD into 50 equally-sized bins. The average ChIP-seq read density (r.p.m./bp) of different factors in each bin was calculated and drawn.

Heatmap Representation of High-confidence ChIA-PET Interactions Super-enhancer Domains and Polycomb-repressed Domains

Heatmap representations of ChIA-PET interactions were created by mapping high-confidence ChIA-PET interactions across Super-enhancer Domains (SD) and Polycomb-repressed Domains (PD), which are defined above. We created three types of regions: upstream, SD or PD, and downstream. Upstream and downstream regions are 20% of the SD's or PD's length each. We divided the

upstream and downstream regions into 10 equally-sized bins each. We divided the SD or PD into 50 equally-sized bins.

To calculate interactions in each bin, we filtered high-confidence in two ways.

1) We required high-confidence interactions to have at least one end in the interrogated region. This removed interactions that are anchored outside of our region of interest.
2) We removed interactions that are not related to the internal structure of the domain. This removed interactions that have one end at an SD or PD border PET peak and the other end outside of the SD or PD.

The density of the whole spans of ChIA-PET interactions in each bin was next calculated in the units of number of interactions per bin. The density of ChIA-PET interactions was row-normalized to the row maximum for each domain and was displayed in Figure 3D and 5D.

Definition of Putative Chromatin Insulator Elements at the Boundaries of Polycomb Domains

An entropy-based measure of Jensen-Shannon Divergence (JSD) was adopted to identify putative SMC1- and CTCF-bound chromatin insulator elements at PD domain boundaries. We divided 20 kb regions centered on CTCF-enriched regions within SDs or PDs into 100 equally-sized bins. We used H3K27me3 and SUZ12 ChIP-seq profiles to identify putative insulator elements at PD boundaries. For each 20 kb region, the average ChIP-seq read density within each bin was calculated and this vector was normalized to the sum of

average read densities so the new normalized vector sums to 1. Since we expect high ChIP-seq signal at one side of insulator elements and low ChIP-seq signal at other side of insulator elements, we defined two vectors to represent the chromatin patterns at insulator elements at the left or right borders of PDs: one vector has 50 0s followed by 50 1s, and the other has 50 1s followed by 50 0s. These vectors were normalized so their sum was 1.

We next used JSD as described in (Fuglede and Topsoe 2004) to quantify the similarity between normalized ChIP-seq patterns and the two pre-defined patterns, which results in a similarity score between each normalized ChIP-seq vector and the ideal vectors described above. We took the top 15 percent of our 20 kb regions ranked by their similarity score and extracted those that were at the boundaries of Polycomb Domains (PD). For robustness, only PD border regions whose average ChIP-seq signal (H3K27me3) within the 20 kb window was above the 60 percentile of all CTCF enriched regions at the side within the domain and below 50 percentile of all CTCF enriched regions at the side outside of the domain were considered as putative chromatin insulator elements. Figure 5E show normalized ChIP-seq density at these putative chromatin insulator elements by standard Z-transform across all CTCF enriched regions.

Conservation of CTCF Binding Across Cell Types

CTCF peaks in 18 tissues/cell types from ENCODE were downloaded from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeLicrTfbs). We restricted our analysis to

autosomal CTCF sites, because these 18 cell types could be derived from mice of different sex or strains. We first took the intersection of autosomal CTCF peaks between our CTCF peaks in murine ESCs and CTCF peaks in the murine ESC Bruce4 line from ENCODE to account for differences in cells and experimental technique. We next quantified how frequently these autosomal CTCF peaks from ESCs were occupied by CTCFs in 18 tissues/cell types (including ESC Bruce4 cells) from ENCODE. The histogram of CTCF occupancy across 18 tissues/cell types were plotted in Figure 6C.

Super-enhancers in NPCs

Super-enhancers were identified in mouse neural progenitor cells (NPCs) using ROSE (https://bitbucket.org/young_computation/rose). This code is an implementation of the method used in (Hnisz et al. 2013, Loven et al. 2013). Briefly, regions enriched in H3K27ac signal were identified using MACS with background control, --keep-dup=auto, and -p 1e-9. These regions were stitched together if they were within 12.5 kb of each other and enriched regions entirely contained within +/- 2 kb from a TSS were excluded from stitching. Stitched regions were ranked by H3K27ac signal therein. ROSE identified a point at which the two classes of enhancers were separable. Those stitched enhancers falling above this threshold were considered super-enhancers.

5C CTCF-CTCF interactions in NPCs

Phillips-Cremins et al. performed 5C at 7 genomic loci (Phillips-Cremins et al. 2013). We filtered for statistically significant 5C interactions in mouse NPC by requiring a p value for both replicates < 0.05, resulting in 674 interactions. We filtered for CTCF-CTCF interactions by requiring an overlap with a CTCF ChIP-Seq enriched region in NPC on both end resulting in 32 CTCF-positive 5C interactions. 34% (11/32) CTCF 5C interactions in NPCs have an 80% reciprocal overlap with a SMC1 ChIA-PET interactions in mouse ESCs (Table S12).

Accession numbers

The GEO accession ID for aligned and raw data is GSE57913 (www.ncbi.nlm.nih.gov/geo/).

SUPPLEMENTAL REFERENCES

Benjamini, Y., Hochberg, Y. 1995. "Controlling the false discovery rate: a practical and pwerful approach to multiple testing." *J. R. Statist. Soc. B* 57 (1):280-300.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, G. W. Bell, A. P. Otte, M. Vidal, D. K. Gifford, R. A. Young, and R. Jaenisch. 2006. "Polycomb complexes repress developmental regulators in murine embryonic stem cells." *Nature* 441 (7091):349-53. doi: nature04733 [pii] 10.1038/nature04733.

Chen, X., H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. 2008. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell* 133 (6):1106-17. doi: S0092-8674(08)00617-X [pii] 10.1016/j.cell.2008.04.043.

Chepelev, I., G. Wei, D. Wangsa, Q. Tang, and K. Zhao. 2012. "Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization." *Cell Res* 22 (3):490-503. doi: 10.1038/cr.2012.15.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. 2012. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485 (7398):376-80. doi: nature11082 [pii] 10.1038/nature11082.

Fuglede, B., and F. Topsoe. 2004. "Jensen-Shannon Divergence and Hilbert space embedding." *Information theory*:31.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. 2009. "An oestrogen-receptor-alpha-bound human chromatin interactome." *Nature* 462 (7269):58-64. doi: nature08497 [pii] 10.1038/nature08497.

Goh, Y., M. J. Fullwood, H. M. Poh, S. Q. Peh, C. T. Ong, J. Zhang, X. Ruan, and Y. Ruan. 2012. "Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation." *J Vis Exp* (62). doi: 10.3791/3770.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. 2007. "A chromatin landmark and transcription initiation at most promoters in human

cells." *Cell* 130 (1):77-88. doi: S0092-8674(07)00681-2 [pii] 10.1016/j.cell.2007.05.042.

Handoko, L., H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. 2011. "CTCF-mediated functional chromatin interactome in pluripotent cells." *Nat Genet* 43 (7):630-8. doi: ng.857 [pii] 10.1038/ng.857.

Hnisz, D, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. 2013. "Transcriptional super-enhancers connected to cell identity and disease." *Cell* In Press.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. 2010. "Mediator and cohesin connect gene expression and chromatin architecture." *Nature* 467 (7314):430-5. doi: 10.1038/nature09380.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. "The human genome browser at UCSC." *Genome Res* 12 (6):996-1006. doi: 10.1101/gr.229102. Article published online before print in May 2002.

Kieffer-Kwon, K. R., Z. Tang, E. Mathe, J. Qian, M. H. Sung, G. Li, W. Resch, S. Baek, N. Pruett, L. Grontved, L. Vian, S. Nelson, H. Zare, O. Hakim, D. Reyon, A. Yamane, H. Nakahashi, A. L. Kovalchuk, J. Zou, J. K. Joung, V. Sartorelli, C. L. Wei, X. Ruan, G. L. Hager, Y. Ruan, and R. Casellas. 2013. "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation." *Cell* 155 (7):1507-20. doi: 10.1016/j.cell.2013.11.039.

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

Lee, T. I., R. G. Jenner, L. A. Boyer, M. G. Guenther, S. S. Levine, R. M. Kumar, B. Chevalier, S. E. Johnstone, M. F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H. L. Murray, J. P. Zucker, B. Yuan, G. W. Bell, E. Herbolsheimer, N. M. Hannett, K. Sun, D. T. Odom, A. P. Otte, T. L. Volkert, D. P. Bartel, D. A. Melton, D. K. Gifford, R. Jaenisch, and R. A. Young. 2006. "Control of developmental regulators by Polycomb in human embryonic stem cells." *Cell* 125 (2):301-13. doi: S0092-8674(06)00384-9 [pii] 10.1016/j.cell.2006.02.043.

Lehoczky, J. A., M. E. Williams, and J. W. Innis. 2004. "Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication." *Evol Dev* 6 (6):423-30. doi: 10.1111/j.1525-142X.2004.04050.x.

Lettice, L. A., S. J. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. 2003. "A long-range Shh enhancer

regulates expression in the developing limb and fin and is associated with preaxial polydactyly." *Hum Mol Genet* 12 (14):1725-35.

Li, G., M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H. S. Ooi, C. Tennakoon, C. L. Wei, Y. Ruan, and W. K. Sung. 2010. "ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing." *Genome Biol* 11 (2):R22. doi: 10.1186/gb-2010-11-2-r22.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. 2012. "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation." *Cell* 148 (1-2):84-98. doi: S0092-8674(11)01517-0 [pii] 10.1016/j.cell.2011.12.014.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. 2009. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326 (5950):289-93. doi: 326/5950/289 [pii] 10.1126/science.1181369.

Loven, J., H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young. 2013. "Selective inhibition of tumor oncogenes by disruption of super-enhancers." *Cell* 153 (2):320-34. doi: S0092-8674(13)00393-0 [pii] 10.1016/j.cell.2013.03.036.

Meuleman, W., D. Peric-Hupkes, J. Kind, J. B. Beaudry, L. Pagie, M. Kellis, M. Reinders, L. Wessels, and B. van Steensel. 2013. "Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence." *Genome Res* 23 (2):270-80. doi: 10.1101/gr.141028.112.

Phillips-Cremins, J. E., M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces. 2013. "Architectural protein subclasses shape 3D organization of genomes during lineage commitment." *Cell* 153 (6):1281-95. doi: S0092-8674(13)00529-1 [pii] 10.1016/j.cell.2013.04.053.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. 2012. "A map of the cis-regulatory sequences in the mouse genome." *Nature* 488 (7409):116-20. doi: nature11243 [pii] 10.1038/nature11243.

Spitz, F., F. Gonzalez, and D. Duboule. 2003. "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." *Cell* 113 (3):405-17.

Wen, B., H. Wu, Y. Shinkai, R. A. Irizarry, and A. P. Feinberg. 2009. "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." *Nat Genet* 41 (2):246-50. doi: 10.1038/ng.297.

Whyte, W. A., S. Bilodeau, D. A. Orlando, H. A. Hoke, G. M. Frampton, C. T. Foster, S. M. Cowley, and R. A. Young. 2012. "Enhancer decommissioning by LSD1 during embryonic stem cell differentiation." *Nature*. doi: nature10805 [pii] 10.1038/nature10805.

Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. 2013. "Master transcription factors and mediator establish super-enhancers at key cell identity genes." *Cell* 153 (2):307-19. doi: 10.1016/j.cell.2013.03.035.

Xu, Z., G. Wei, I. Chepelev, K. Zhao, and G. Felsenfeld. 2011. "Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription." *Nat Struct Mol Biol* 18 (3):372-8. doi: 10.1038/nsmb.1993.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol* 9 (9):R137. doi: 10.1186/gb-2008-9-9-r137.

Zhang, Y., C. H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, F. H. Mulawadi, W. K. Sung, S. Nicolis, N. Ahituv, Y. Ruan, and C. L. Wei. 2013. "Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations." *Nature* 504 (7479):306-10. doi: 10.1038/nature12716.