

# **Study of the human *SOX17* locus and its genetic determinants in definitive endoderm**

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

MTLA B.Sc. M.Sc. Alexandro Landshammer

from Ulm /Donau

2022

I conducted my doctoral studies from June 2018 until June 2022 at the Max Planck Institute for Molecular Genetics, Berlin Germany under the guidance of Prof. Dr. Alexander Meissner.

1<sup>st</sup> Reviewer: Prof. Dr. Alexander Meissner  
Max-Planck-Institute for Molecular Genetics, Berlin Germany

2<sup>nd</sup> Reviewer: Prof. Dr. Katja Nowick  
Institute of Biology – Zoology, Freie Universität Berlin Germany

Date of Disputation: 22.08.2022

# TABLE OF CONTENT

ABSTRACT .....	IV
ZUSAMMENFASSUNG.....	V
LIST OF PERSONAL PUBLICATIONS AND CONTRIBUTIONS .....	VII
OTHER PERSONAL CONTRIBUTIONS .....	VII
LIST OF ABBREVIATIONS.....	VIII
LIST OF FIGURES.....	XI
1 INTRODUCTION .....	1
1.1 Human gastrulation and the development of definitive endoderm .....	1
1.1.1 The formation of definitive endoderm .....	1
1.1.2 <i>In vitro</i> definitive endoderm derivation as a model system .....	3
1.1.3 Phylogeny and function of SOX genes .....	5
1.2 Gene regulation in development .....	8
1.2.1 The non-coding genome: <i>cis</i> -regulatory elements (CREs) .....	8
1.2.2 Promoters and Enhancers .....	9
1.2.3 Regulation by distal enhancers.....	11
1.3 The 3-dimensional (3D) genome.....	13
1.3.1 Chromosome compartments.....	14
1.3.2 Topologically associating domains (TADs).....	14
1.3.3 Intra-TAD interaction dynamics.....	15
1.3.4 The role of CTCF in 3D chromatin organization .....	16
1.4 Non-coding RNAs (ncRNAs) .....	17
1.4.1 Enhancer RNAs (eRNAs).....	18
1.4.2 Long non-coding RNAs (lncRNAs) .....	18
1.5 The human <i>SOX17</i> locus in definitive endoderm .....	21
1.5.1 Definitive endoderm and its DNA methylome.....	21
1.5.2 <i>SOX17</i> and its diverse functions in development and disease .....	23
2 AIMS OF THE THESIS.....	26

3	THESIS CONTRIBUTIONS.....	27
3.1	Publication: Topological isolation of developmental regulators in mammalian genomes. .....	27
3.2	Unpublished: Discovery and characterization of <i>LNC</i> <i>SOX17</i> as an essential regulator in endoderm formation. ....	28
4	RESULTS.....	29
4.1	Identification of single gene loop domains and topologically isolated genes (TIGs) ....	29
4.2	Characterization of CTCF loop domains and their boundary elements .....	32
4.3	Genetic dissection and characterization of the <i>SOX17</i> loop domain .....	36
4.4	Epigenetic profiling of the <i>SOX17</i> -DMR and identification of a novel lncRNA locus ....	45
4.5	Characterization of the novel long non-coding RNA (lncRNA) <i>LNC</i> <i>SOX17</i> .....	49
4.6	<i>LNC</i> <i>SOX17</i> does not regulate <i>SOX17</i> in <i>cis</i> during definitive endoderm .....	53
4.7	Lack of <i>LNC</i> <i>SOX17</i> leads to aberrant definitive endoderm and differentiation failure .	57
5	DISCUSSION.....	62
5.1	Topological isolation and its regulatory importance for developmental genes as <i>SOX17</i> .....	63
5.2	The <i>SOX17</i> DMR is comprised of two distinct CREs of diverse function .....	67
5.3	<i>LNC</i> <i>SOX17</i> is dependent on <i>SOX17</i> and does not regulate its locus in <i>cis</i> .....	70
5.4	Loss of <i>LNC</i> <i>SOX17</i> leads to an aberrant definitive endoderm phenotype .....	74
5.5	Conclusion .....	78
6	REFERENCES .....	80
7	MATERIALS AND METHODS.....	98
7.1	Experimental Materials and Approaches .....	98
7.1.1	Molecular cloning.....	98
7.1.2	Sequencing and library preparations.....	105
7.1.3	Cell culture and generation of transgenic/targeted cell lines .....	111
7.1.4	Differentiation and inhibitor assays .....	117
7.1.5	Imaging/FACS based assays .....	119
7.1.6	PCR based assays .....	127

7.1.7 Miscellaneous assay types.....	134
7.2 Computational Methods and Approaches.....	137
7.2.1 Parameters .....	137
7.2.2 Identification of CTCF loop domains from Hi-C data .....	137
7.2.3 Clustering and merging of redundant loops .....	137
7.2.4 Boundary anchored virtual 4C visualization of Hi-C data .....	138
7.2.5 Identification of CTCF motifs and their conservation across species.....	138
7.2.6 Evolutionary analysis of human CTCF loop domain boundaries.....	138
7.2.7 Identification of consensus CTCF binding sites .....	138
7.2.8 Clustered and typical CTCF-binding sites.....	139
7.2.9 Identification of enhancers and analysis of their H3K27ac enrichment.....	139
7.2.10 Gene sets and enrichment analysis .....	140
7.2.11 Chromatin confirmation capture sequencing analysis .....	140
7.2.12 Coding potential calculation.....	140
7.2.13 ChIP sequencing .....	141
7.2.14 Single-cell RNAseq pipeline .....	141
7.2.15 Bulk measurements from scRNAseq pipeline.....	141
7.2.16 Oxford Nanopore RNA analysis .....	142
7.2.17 Oxford Nanopore RNA split-read analysis .....	142
7.2.18 RNA sequencing data analysis .....	143
7.2.19 Data visualization.....	143
7.2.20 Data and code availability.....	144
8 APPENDICES .....	145
8.1 Acknowledgement .....	145
8.2 Declaration of Academic Integrity .....	146
8.3 Curriculum Vitae .....	147
8.4 Attachment.....	150

## ABSTRACT

Embryonic development and organogenesis depend on the precise spatiotemporal expression of specific sets of genes. Precisely controlled gene expression ensures cell state transitions, especially in the early stages of development, as gastrulation. These complex multi-layered cellular processes are orchestrated by the interfacing of the epigenome, 3-dimensional (3D) nuclear organization, *cis*-regulatory elements (CREs) with transcription factors (TF), and long non-coding RNAs (lncRNAs). In the gastrulating embryo, definitive endoderm is specified from the pluripotent epiblast following a series of regulatory events, including the activation of SOX17, a key TF of that particular germ layer. Although SOX17 has been extensively studied in early embryonic development, the precise control of its activation, the locus, and the epigenetic rules governing its genetic regulatory network (GRN) remains poorly investigated. In my thesis, I in-depth characterized the human *SOX17* locus, exploring the relevance and regulatory impact of 3D nuclear organization, its distal CREs, and their activity. I applied a series of loss of function (LOF) and transgenic experiments to dissect the locus at a satisfactory resolution. In particular, I showed *SOX17* among a subset of developmental regulators topologically isolated within CTCF-CTCF loop domains and highlighted the importance of gene control in 3D within this type of domain. I pinpointed the relevance of *SOX17*'s distal CREs and their definitive endoderm-specific interaction and showed this interaction to be highly dependent on CTCF-CTCF loop-formation to guarantee proper gene control. I found CRE-dependent *SOX17* gene deregulation associated with poor definitive endoderm differentiation outcome and a stalled "mesendodermal-like" phenotype. Assessing the genetic identity of different CREs, I divulged the presence of a novel lncRNA within the locus, namely *LNC*SOX17. I fully characterized *LNC*SOX17 and established its identity as a *bona fide* lncRNA through a series of genetic perturbations. I demonstrated the importance of *LNC*SOX17 for forming definitive endoderm and the lack of participation in *SOX17 cis*-acting gene control. I associated the loss of *LNC*SOX17 RNA but not its active transcription at the locus with an aberrant endodermal transcriptome, a lack of epithelial-to-mesenchymal transition (EMT), and the hyperactivity of the detrimental definitive endoderm JNK/JUN/AP1 signaling pathway. I found definitive endoderm lacking *LNC*SOX17 to be functionally impeded in the generation of pancreatic progenitor populations. The studies within this thesis serve as valuable examples to support the functional relevance of 3D nuclear organization and its importance for developmental gene control in *cis* via CTCF-CTCF loop domain-mediated CRE-promoter contact facilitation. They associate developmental gene expression levels with various phenotypes, identify a so far unknown developmental lncRNA molecule, and imply its relevance for the formation of definitive endoderm. The outlined results advance our knowledge of developmental TF gene-control and its importance for the development of human definitive endoderm.

## ZUSAMMENFASSUNG

Embryonale Entwicklung und Organogenese sind abhängig von präziser raumzeitlicher Expression einer Reihe spezifischer Gene. Präzise kontrollierte Genexpression, garantiert Zellzustandsübergänge, insbesondere in den frühen Stadien der Entwicklung, wie der Gastrulation. Diese komplexen, vielschichtigen Prozesse, werden durch die Verbindung des Epigenoms, 3-dimensionaler (3D) nukleärer Anordnung, *cis*-regulatorischen Elementen (CRE) mit Transkriptionsfaktoren (TF) und langen nicht-kodierenden RNAs (lncRNAs) instrumentiert. Im gastrulierenden Embryo, spezifiziert sich Endoderm ausgehend vom pluripotenten Epiblast, mit einer darauffolgenden Serie an regulatorischen Ereignissen, einschließlich der Aktivierung von SOX17, ein Schlüssel-TF dieses speziellen Keimblatts. Obwohl SOX17 im Kontext von embryonaler Frühentwicklung bisher ausgiebig studiert wurde, sind die präzise Kontrolle seiner Aktivierung, sein Genort und die epigenetischen Regulatoren die sein genregulatorisches Netzwerk (GRN) steuern, immernoch unzureichend untersucht. In meiner Dissertation charakterisierte ich ausführlich den menschlichen Genort SOX17, und erforschte die Relevanz und den regulatorischen Einfluss 3D nukleärer Anordnung, distale CREs und deren Aktivität. Hierfür wandte ich eine Serie von Funktionsverlust- (LOF) und Transgenexperimenten an, um den Genort in ausreichender Auflösung zu sezieren. Im Besonderen konnte ich zeigen, dass SOX17 einer Subgruppe von Genen angehört, die topologisch innerhalb von CTCF-CTCF Schleifendomänen isoliert sind und hob die Tragweite von Genkontrolle in 3D, innerhalb dieses Domäentyps hervor. Des Weiteren bestimmte ich die Relevanz distaler CREs und deren endoderm-spezifische Interaktion mit SOX17, und zeigte auf, dass diese Interaktion höchst abhängig von CTCF-CTCF Schleifenbildung für die Gewährleistung korrekter Genkontrolle ist. Ich war in der Lage CRE-abhängige, SOX17-Genregulation mit verschlechterter endodermaler Differenzierung, im Sinne eines „mesendodermal-ähnlichen“ Sackgassenphänotyps zu assoziieren. Im Zuge der genetische Identitätsbeurteilung verschiedener CREs, enthüllte ich das Auftreten einer bis dato unbekanntem Lokusassoziierten, neuen lncRNA, nämlich LNC SOX17. Durch eine Reihe genetischer Manipulationen, konnte ich LNC SOX17 vollständig charakterisieren und bewies seine Identitätsechtheit als lncRNA. Des Weiteren war ich in der Lage die Wichtigkeit LNC SOX17s für die Endodermaleentwicklung zu zeigen und entkräftete ihr Einwirken an der *cis*-abhängigen Genregulation von SOX17. Des Weiteren konnte ich den Verlust von LNC SOX17 – und nicht dessen Transkription am Genort – mit veränderter, endodermaler Genexpression, einem Ausbleiben von Epithelialer-zu-Mesenchymaler Transition (EMT) und der Hyperaktivität des endodermbeeinträchtigendem JNK/JUN/AP1 Signalwegs assoziieren. Zudem entdeckte ich, dass Endoderm ohne LNC SOX17, funktional in der Weiterentwicklung zu pankreatischem Vorläufergewebe eingeschränkt ist. Die Studien dieser Arbeit dienen als wertvolle Beispiele zur Unterstützung funktioneller Relevanz von 3D nukleärer Anordnung, und

deren Wichtigkeit für Genregulation während der Frühentwicklung in *cis*, durch CTCF-Schleifendomänen vermittelter CRE-Promoter Kontakterleichterung. Sie assoziieren Genexpression mit verschiedenen Phänotypen, während der endodermalen Frühentwicklung, zeigen die Identifizierung eines bis dato noch unbekanntes lncRNA-Moleküls auf, und weisen auf dessen Relevanz für die korrekte Ausbildung des Endoderms hin. Durch die Ergebnisse dieser Arbeit erweiterten wir unser Wissen im Bezug auf TF-Genkontrolle während embryonaler Frühentwicklung und dessen Wichtigkeit für die Entwicklung des menschlichen Endoderms.



## LIST OF PERSONAL PUBLICATIONS AND CONTRIBUTIONS

Hua-Jun Wu\*, **Alexandro Landshammer\***, K. Stamenova, Adriano Bolondi, Helene Kretzmer, Alexander Meissner & Franziska Michor. *Nature Communications* (2021) 12:4897. doi: 10.1038/s41467-021-24951-7. Topological isolation of developmental regulators in mammalian genomes.

\* Wu and Landshammer contributed equally to the study

## OTHER PERSONAL CONTRIBUTIONS

Hazel M Quinn, Regina Vogel, Oliver Popp, Philipp Mertins, Linxiang Lan, Clemens Messerschmidt, **Alexandro Landshammer**, Kamil Lisek, Sophie Château-Joubert, Elisabetta Marangoni, Elle Koren, Yaron Fuchs, Walter Birchmeier. *Cancer Research* (2020) 15;81(8):2116-2127. doi: 10.1158/0008-5472.CAN-20-2801. YAP and  $\beta$ -catenin cooperate to drive oncogenesis in basal breast cancer.

Rashmi Tandon, Bjoern Braendl, Natalya Baryshnikova, **Alexandro Landshammer**, Laura Steenpaß, Oliver Keminer, Ole Pless, Franz-Josef Müller. *Stem Cell Research* (2018) 33:120-124. doi: 10.1016/j.scr.2018.10.004. Generation of two human isogenic iPSC lines from fetal dermal fibroblasts.

Lotta von Boehmer\*, Muriel Mattle\*, Peter Bode, **Alexandro Landshammer**, Carolin Schaefer, Natko Nuber, Gerd Ritter, Lloyd Old, Holger Moch, Niklaus Schaefer, Elke Jaeger, Alexander Knuth, Maries van den Broek. *Cancer Immunology* (2013) 13:12. Print 2013. PMID: PMC3718732. NY-ESO-1-specific immunological pressure and escape in a patient with metastatic melanoma.

\* von Boehmer and Mattle contributed equally to the study

Anurag Gupta, Natko Nuber, Christoph Esslinger, Mareike Wittenbrink, Martin Treder, **Alexandro Landshammer**, Takuro Noguchi, Marcus Kelly, Sacha Gnjatic, Erika Ritter, Lotta von Boehmer, Hiroyoshi Nishikawa, Hiroshi Shiku, Lloyd Old, Gerd Ritter, Alexander Knuth, Maries van den Broek. *Cancer Immunology* (2013) 13: 3. PMID: PMC3559191. A novel human-derived antibody against NY-ESO-1 improves the efficacy of chemotherapy.

Anurag Gupta, Hans Christian Probst, Van Vuong, **Alexandro Landshammer**, Sabine Muth, Hideo Yagita, Reto Schwendener, Martin Pruschy, Alexander Knuth, Maries van den Broek. *Journal of Immunology* (2012) 189(2):558-66. doi: 10.4049/jimmunol.1200563. Radiotherapy promotes tumor-specific effector CD8+ T cells via dendritic cell activation.

## LIST OF ABBREVIATIONS

### #

3C	chromatin confirmation capture
4C	circular chromatin confirmation capture
3D	3-dimensional
5mC	5-methylcytosine

### B

bp	base pair
----	-----------

### C

CAM	cell adhesion molecule
cHi-C	capture Hi-C
ChIP	chromatin immunoprecipitation
cic	capicua
CpG	CpG dinucleotide motif
CRE	<i>cis</i> -regulatory element
CRISPR	clustered regularly interspaced short palindromic repeats
CRISPRi	CRISPR interference
CTCF	CCCTC-binding factor
CXCR4	C-X-C chemokine receptor 4

### D

DI	directionality index
DEG	differentially expressed gene
DNA	deoxyribonucleic acid
DMR	differentially methylated region
DR	developmental regulator
DRE	distal regulatory element

### E

EC	ectoderm
eDR	early developmental regulator
EN	definitive endoderm, endoderm
EMT	epithelial to mesenchymal transition
eRNA	enhancer RNA
ESC	embryonic stem cell

### F

FISH	fluorescent <i>in situ</i> hybridization
FACS	fluorescent activated cell sorting

**G**

GSEA	gene set enrichment analysis
GRN	gene(tic) regulatory network
GO	gene ontology

**H**

HMG	high-mobility-group
Hi-C	genome wide chromosome conformation capture
hPSC	human pluripotent stem cells

**I**

iPSC	induced pluripotent stem cell
------	-------------------------------

**K**

kb	kilo base-pair
----	----------------

**L**

lincRNA	long intergenic non-coding RNA
LNA	locked nucleic acid
lncRNA	long non-coding RNA

**M**

m <sup>7</sup> G	7-methyl guanosine
ME	mesoderm
MET	mesenchymal to epithelial transition

**N**

n.a.	not available
ncDNA	non-coding DNA = "junk" DNA

**O**

ORF	open-reading-frame
-----	--------------------

**P**

PCA	principal component analysis
PCG	protein-coding gene
PRC	polycomb repressive complex
PSC	pluripotent stem cell

**Q**

qRT-PCR	quantitative real-time polymerase chain-reaction
---------	--

**R**

RACE	rapid amplification of cDNA ends
RBD	RNA-binding domain
RBP	RNA binding protein
RNA	ribonucleic acid
RNAi	RNA interference

**S**

scRNA	single-cell RNA
shRNA	small hairpin RNA
SOX	<i>SRY</i> -related high-mobility-group-box
SRY	sex determining region of Y-gen
SE	super-enhancer

**T**

TAD	topologically associating domain
TF	transcription factor
TGF- $\beta$	transforming growth factor- $\beta$
TSS	transcriptional start site

**W**

WNT	wingless / integrated
-----	-----------------------

**X**

X-LAG	X-linked acrogigantism
-------	------------------------

**Z**

ZGA	zygotic genome activation
-----	---------------------------

## LIST OF FIGURES

<b>Fig. 1</b>	Human gastrulation and the formation of the three germ-layer	<b>2</b>
<b>Fig. 2</b>	Human gastrulation: Key marker of EMT to MET transition	<b>3</b>
<b>Fig. 3</b>	<i>In vitro</i> differentiation into the three germ-layer from human ESC/iPSC	<b>5</b>
<b>Fig. 4</b>	Transcriptional apparatus governed by promoter and enhancer elements	<b>10</b>
<b>Fig. 5</b>	Model for transcriptional condensation and gene control	<b>12</b>
<b>Fig. 6</b>	Scheme of the 3D-genome organizational layers	<b>14</b>
<b>Fig. 7</b>	Modes of lncRNA action	<b>20</b>
<b>Fig. 8</b>	DMRs across the three germ-layers	<b>22</b>
<b>Fig. 9</b>	DNA methylation landscape of the <i>SOX17</i> locus	<b>23</b>
<b>Fig. 10</b>	Identification of topologically insulated CTCF loop domains in HUES64 ESC genomes	<b>29</b>
<b>Fig. 11</b>	CTCF single gene loop domains and their topologically isolated genes (TIGs)	<b>31</b>
<b>Fig. 12</b>	CTCF loop domain establishment and stability throughout early development	<b>32</b>
<b>Fig. 13</b>	Gene ontology (GO), conservation, and phylogeny of CTCF loop domains	<b>34</b>
<b>Fig. 14</b>	CTCF loop domain strength and CRE abundance	<b>35</b>
<b>Fig. 15</b>	Overview of the <i>SOX17</i> CTCF loop domain at a locus resolution	<b>36</b>
<b>Fig. 16</b>	Boundary 2 CRISPR/Cas9 perturbation of the <i>SOX17</i> CTCF loop domain in detail	<b>38</b>
<b>Fig. 17</b>	<i>SOX17</i> Boundary 2 perturbation leads <i>SOX17</i> deregulation without affecting local gene expression	<b>40</b>
<b>Fig. 18</b>	Comprehensive gene expression analysis of wildtype and Boundary 2 perturbed endoderm	<b>41</b>
<b>Fig. 19</b>	<i>SOX17</i> Boundary 2 caused gene-deregulation leads to a “mesendodermal like” state	<b>42</b>
<b>Fig. 20</b>	<i>SOX17</i> Boundary 2 perturbation rescue experiment	<b>44</b>
<b>Fig. 21</b>	Epigenetic profiling and genetic characterization of the <i>SOX17</i> -DMR	<b>46</b>
<b>Fig. 22</b>	Identification of the novel lncRNA locus <i>LNC<sup>SOX17</sup></i> and functional validation of <i>pLNC<sup>SOX17</sup></i>	<b>48</b>
<b>Fig. 23</b>	<i>LNC<sup>SOX17</sup></i> shows high definitive endoderm specificity and developmental conservation	<b>50</b>

<b>Fig. 24</b> <i>LNC</i> <i>SOX17</i> is a <i>SOX17</i> uncoupled RNA with partially defined processing, start and end	<b>51</b>
<b>Fig. 25</b> <i>LNC</i> <i>SOX17</i> is a nuclear long non-coding RNA	<b>52</b>
<b>Fig. 26</b> CRISPR interference (CRISPRi) based repression of <i>LNC</i> <i>SOX17</i>	<b>53</b>
<b>Fig. 27</b> CRISPRi repression of <i>LNC</i> <i>SOX17</i> does not influence <i>SOX17</i> gene control in <i>cis</i>	<b>54</b>
<b>Fig. 28</b> CRISPR/Cas9 <i>SOX17</i> perturbation shows <i>SOX17</i> dependence for <i>LNC</i> <i>SOX17</i> expression	<b>55</b>
<b>Fig. 29</b> CRISPR/Cas9 integrated early transcriptional termination phenocopies repression of <i>LNC</i> <i>SOX17</i>	<b>56</b>
<b>Fig. 30</b> Both, absence, and repression of <i>LNC</i> <i>SOX17</i> lead to CXCR4 deregulation	<b>57</b>
<b>Fig. 31</b> Repression of <i>LNC</i> <i>SOX17</i> leads to an endoderm specific aberrant transcriptome	<b>58</b>
<b>Fig. 32</b> Repression of <i>LNC</i> <i>SOX17</i> leads JNK hyperactivity and EMT failure	<b>59</b>
<b>Fig. 33</b> <i>LNC</i> <i>SOX17</i> repressed endoderm lacks the ability to generate PP1 pancreatic progeny	<b>60</b>
<b>Fig. 34</b> <i>Cis</i> - versus <i>trans</i> -regulation of <i>SOX17</i> and its genetic determinants in definitive endoderm	<b>78</b>
<b>Fig. 35</b> Association of genetic dissections at the <i>SOX17</i> locus with definitive endoderm and pancreatic differentiation phenotypes	<b>79</b>

# 1 INTRODUCTION

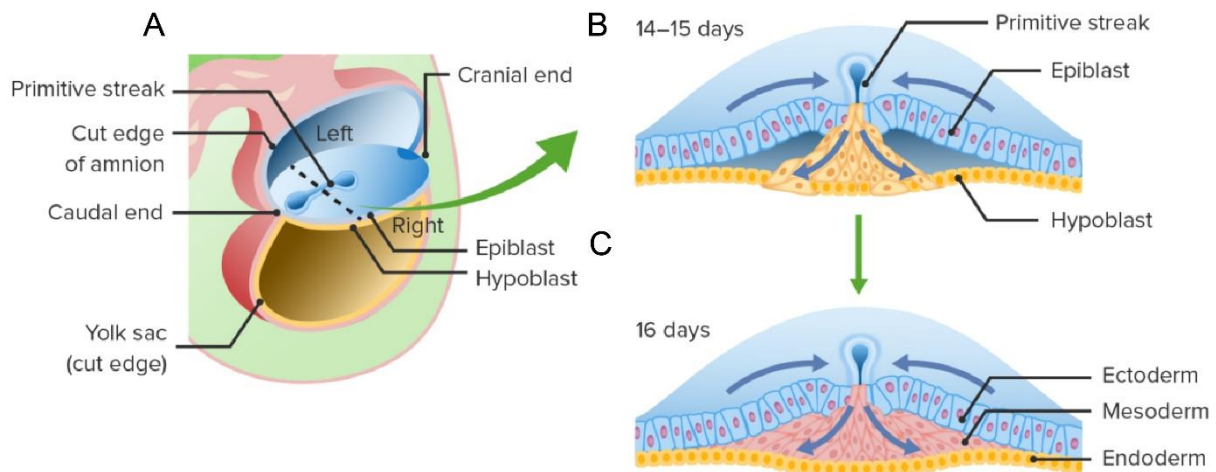
Gastrulation, which is truly the most important time in your life – even more important than birth, marriage, or death[1] – begins with the formation of the primitive streak. It is a region in the early human embryo where epiblast cells converge in a well-defined spatial and temporal sequence to form further the three germ-layers, ectoderm, mesoderm, and definitive endoderm.

## 1.1 Human gastrulation and the development of definitive endoderm

The primitive streak is an accumulation of cells at a linear midline derived from the epiblast at the caudal region of the embryo (Fig. 1A). As epiblast cells reach the primitive streak, they change shape and pass through it on their way to forming new layers beneath the epiblast (Fig. 1B). Induction of primitive streak formation is caused by cells at the edge of the embryonic disk, where transforming growth factor- $\beta$  (TGF- $\beta$ ) and WNT family signaling molecules are released. Upon that induction, cells entering the primitive streak form distinct lineages as they leave. The most caudal cells both to enter and leave the streak as it elongates form the extraembryonic mesoderm lining the trophoblast and yolk sac, as well as that forming the blood islands. Another wave of mesoderm, arising later and more cranial in the primitive streak, forms the paraxial, lateral plate, and cardiac mesoderm. A final wave, which enters and leaves the cranial most end of the primitive streak, gives rise to midline axial structures and the embryonic endoderm (Fig. 1B,C).[2, 3]

### 1.1.1 The formation of definitive endoderm

Starting in early human gastrulation at day 14-15 post-fertilization, epiblast cells produce hyaluronic acid, which enters the space between the epiblast and hypoblast (Fig. 1B).[4]. Hyaluronic acid, a polymer consisting of repeating subunits of *D*-glucuronic acid and *N*-acetylglucosamine, is frequently associated with cell migration in developing systems. The molecule has a tremendous capacity to bind water (up to 1000 times its own volume), and it functions to keep mesenchymal cells from aggregating during cell migrations. Although after leaving the primitive streak, the mesenchymal cells of the embryonic mesendoderm – a progenitor state for mesoderm and endoderm – find themselves in a hyaluronic acid-rich environment. In all vertebrate embryos that have been investigated to date, the spread of mesendodermal cells away from the primitive streak or the equivalent structure is found to depend on the presence of fibronectin associated with the basal lamina beneath the epiblast. The embryonic mesendoderm ultimately spreads laterally as a thin sheet of mesenchymal cells between the epiblast and hypoblast layers (Fig. 1B). When the further mesoderm has formed a discrete layer in the human embryo, the upper germ layer (remains of the former epiblast) is



**Fig. 1 Human gastrulation and the formation of the three germ-layer. (A)** Sagittal plane of the human embryo, indicating the formation of the primitive streak along the primitive groove from caudal to cranial. **(B)** Coronal plane of the embryonic disc at day 14-15 post-fertilization (zoom in dashed line at **(A)**), highlighting the first events of human gastrulation. Epiblast cells displace hypoblast cells and form embryonic definitive endoderm via a transition phase of mesendoderm. **(C)** Coronal plane of the embryonic disc at day 16 post-fertilization, highlighting the final events of human gastrulation. Residual epiblast cells turn into future ectoderm and cells in between ecto- and endoderm become mesoderm. (Adapted and modified figure from <https://www.lecturio.com>)

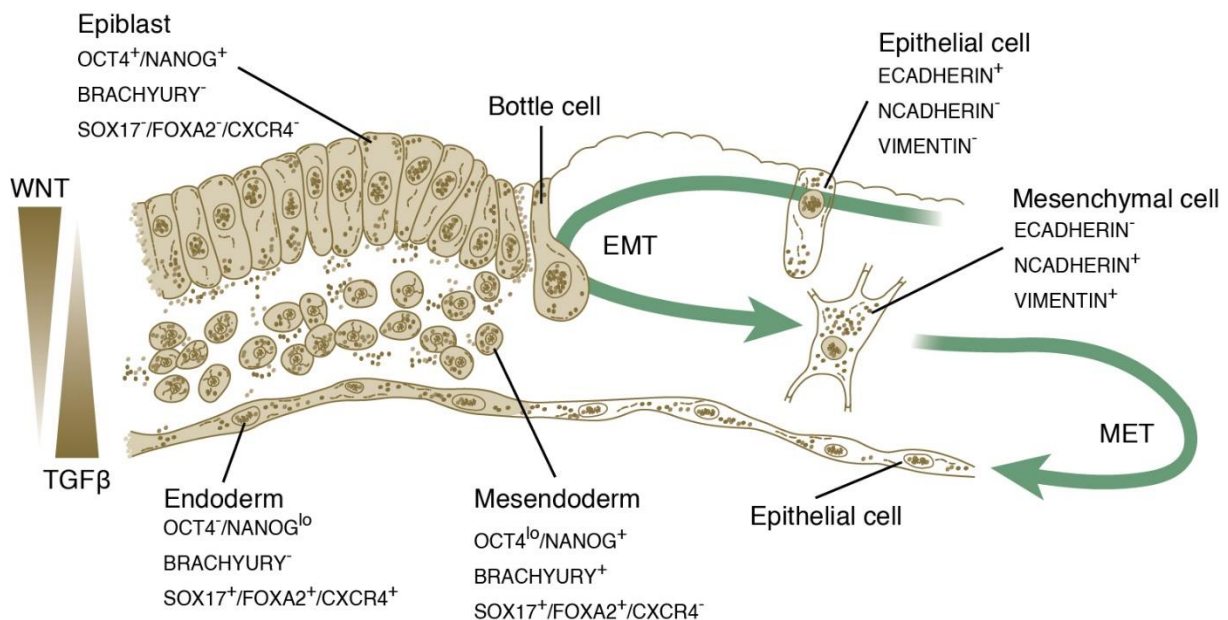
called the ectoderm, and the lower germ layer, which has displaced the original hypoblast, is called the definitive endoderm (Fig. 1C Endoderm).[2, 3]

The movements of the cells passing through the primitive streak are accompanied by major changes in their structure and organization (Fig. 2). While in the epiblast, the cells have the properties of typical epithelial cells, with well-defined apical and basal surfaces, and they are associated with a basal lamina that underlies the epiblast. As they enter the primitive streak, these cells elongate, lose their basal lamina, and take on a characteristic morphology that has led to their being called bottle cells (Fig. 2 center). When they become free of the epiblastic layer in the primitive groove, the bottle cells assume the morphology and characteristics of mesenchymal cells, which can migrate as individual cells if provided with the proper extracellular environment (Fig. 2 right). This transformation includes the loss of specific cell adhesion molecules (CAMs), in particular ECADHERIN (*CDH1*), as the cells convert from an epithelial to a mesenchymal configuration. This transformation is correlated with the expression of the transcription factor (TF) SNAIL (*SNAI1*). As cells in the epiblast are undergoing epithelial-to-mesenchymal transition (EMT), they begin to express NCADHERIN (*CDH2*) and VIMENTIN (*VIM*), which are necessary for their spreading out from the primitive streak in the newly forming mesendodermal layer. This process is reverted as soon as cells replace the original hypoblast converting back from their mesenchymal into the epithelial configuration (Fig. 2 right).[5, 6] Along with these temporally coordinated cellular events, cells of the epiblast also start to differentiate and exit pluripotency which is indicated by the downregulation of pluripotency factors, e.g. OCT4/NANOG (Fig. 2 left) and the upregulation of



mesendodermal marker genes, e.g. BRACHYURY (*T/TBXT*). Further in differentiation, these cells eventually start to co-express definitive endoderm markers, e.g. SOX17/FOXA2. Shortly after inducing endodermal TFs, differentiating cells will lose their mesendodermal character as the cells will downregulate BRACHYURY expression and immediately gain definitive endoderm marker co-expression, e.g. CXCR4 (Fig. 2 left). Most importantly, the expression of CXCR4 is exclusive for definitive and not found in primitive or extraembryonic endodermal tissues, e.g. parietal/visceral endoderm[7].

### 1.1.2 *In vitro* definitive endoderm derivation as a model system



**Fig. 2 Human gastrulation: Key marker of EMT to MET transition.** Highlighted in colored cells on the left is the process of cell-type transitions from pluripotent epiblast cells out of the primitive streak via an intermediate mesendodermal state to the final endoderm cell state along the concentration gradients of TGF-β and Wnt signaling molecules. Hollow cells on the right depict the Epithelial-To-Mesenchymal (EMT) transition process and back (MET). Protein expression levels are depicted below cell-types. (Adapted and modified figure from Carlson B.M. et al, 2014)

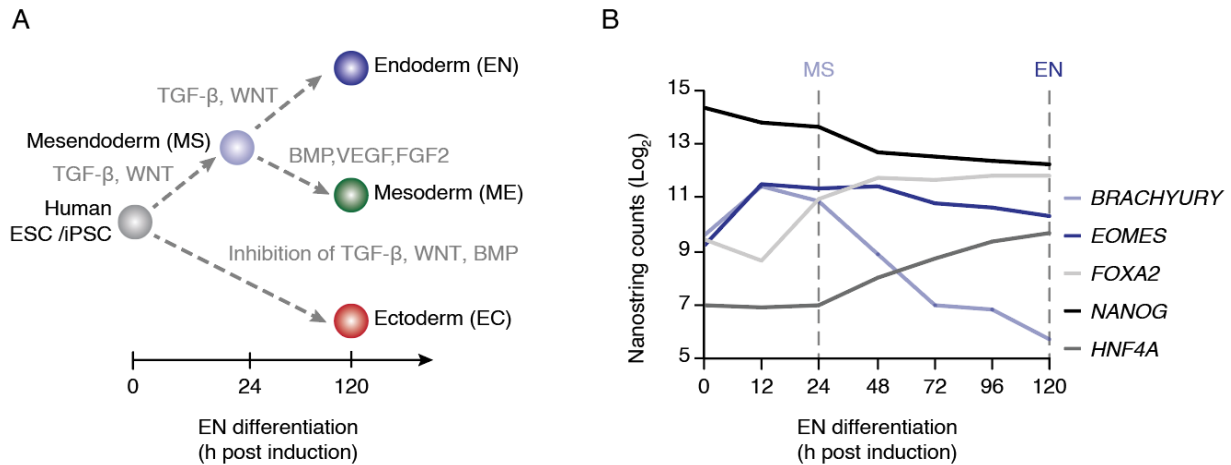
Definitive endoderm is a highly relevant tissue within the embryo since it will give rise to several organs such as the liver and pancreas which are potential targets for cell-based therapy, and so there is great interest in understanding the pathways that regulate the induction and specification of this germ-layer[8].

Similar as in human, mouse embryonic stem cell (mESC) cultures will efficiently induce definitive endoderm upon high levels of Activin/Nodal (TGF-β) signaling (Fig. 2 left)[9, 10]. When analyzed at the primitive streak stage, one step prior to induction of definitive endoderm, Activin-induced mESC populations identified either by the co-expression of Brachyury and Foxa2 or expression of the anterior marker Goosecoid (*Gsc*) were found to contain both mesoderm and endoderm[11, 12]. Clonal analysis revealed that individual cells within the Goosecoid population had the potential to generate both endoderm and mesoderm derivatives,

suggesting that they may represent mesendoderm progenitors (Fig. 2). Thus, the first step in the generation of definitive endoderm may be the formation of mesendodermal progenitors. Progression of the anterior primitive streak population to definitive endoderm depends on sustained Activin signaling[13], consistent with increased Nodal signaling required for definitive endoderm formation in the early mouse embryo[14]. Interestingly, when exposed to high levels of Activin, the *Brachyury*<sup>+</sup>/*Foxa2*<sup>lo</sup> posterior primitive streak population is also able to generate endoderm, indicating that germ-layer fates are not yet fixed at the primitive streak stage in mESC differentiation cultures. Once induced, definitive endoderm forms an epithelial sheet that further undergoes specification to distinct regions known as foregut, midgut, and hindgut[15]. This specification is controlled in part by factors secreted by surrounding mesoderm-derived tissues. As in human, in the gastrulating mouse embryo, *Cxcr4* is expressed in the definitive endoderm but not in primitive endoderm/visceral endoderm[16].

It is important to note that despite morphological differences between human and mouse gastrulation a fundamental transcriptomic similarity has been revealed providing insight to mammalian evolution[17-19]. To study the phenomenon of early human gastrulation *ex utero* and to overcome the limitations of this process taking place post-implantational, it has been put a tremendous amount of effort to gain insights into the cellular and molecular understanding by the development of various 2D[20, 21] and 3D[22, 23] human culture models, in a directed[9, 10, 24, 25] or random[26, 27] differentiation fashion. To do so, it has mostly been utilized human embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) for the derivation of human *in vitro* definitive endoderm. Both cell systems have been shown to share highly equal transcriptional identity[28] and are both equally suitable to generate this tissue derivative *in vitro*.

To understand fate decision-making and to resolve temporal molecular changes during human gastrulation, previous studies have investigated transcriptional dynamics of human ESCs undergoing three germ-layer differentiation, in particular definitive endoderm formation *in vitro*[21, 29]. Doing so, it has been utilized physiological differentiation conditions by high TGF- $\beta$  (ACTIVIN) and low WNT signaling to induce definitive endoderm (Fig. 3A), following downstream transcriptomic analysis (Fig. 3B). *FOXA2* and *SOX17* (not shown) – which only get expressed after 24h of differentiation – highlight the timepoint of arising progenitors capable of giving rise to *in vitro* derived definitive endoderm and mesoderm by co-expression of *BRACHYURY*, earlier referred as mesendoderm. Late endodermal marker genes as e.g. *HNF4A* get expressed only after 24h post-induction together with sustained *FOXA2* and *EOMES* expression levels (Fig. 3B). At this stage continuing stimulation with high levels of



**Fig. 3** *In vitro* differentiation into the three germ-layer from human ESC/iPSC. **(A)** Temporal treatment scheme of the three germ-layer. Endoderm (dark blue) and Mesoderm (green) derive from common Mesendoderm (light blue) progeny. **(B)** Temporal RNA expression scheme of derived endoderm highlighting endodermal master regulator and upstream marker gene *EOMES* (dark blue) and mesendodermal marker gene *BRACHYURY* (light blue). Note the downregulation of *BRACHYURY* (light blue) after 24h and the increase of *FOXA2* (light grey) after passing mesendoderm. Also note the temporally delayed upregulation of endodermal marker *HNF4A* (dark grey), the maintenance of *EOMES* (dark blue) and the overall decrease of pluripotency marker *NANOG* (black) over the course of endoderm formation. (Adapted and modified figure from Tsankov A. et al., 2015)

TGF- $\beta$  (ACTIVIN) and low WNT signaling leads *BRACHYURY* expression levels to drop further, accompanied by an overall expression decrease of pluripotency factor *NANOG*, indicating proper exit of pluripotency and the formation of *in vitro* derived definitive endoderm within 5 days (Fig. 3B). Overall, as *in vivo*, *in vitro* derived definitive endoderm is characterized by decreased expression levels of the pluripotency marker *NANOG* (Fig. 2 and Fig. 3B) and absence of mesendodermal marker *BRACHYURY* (Fig. 3B). Moreover, *in vitro* derived definitive endoderm is also expressing high levels of *FOXA2* and *SOX17/CXCR4* (not shown) (Fig. 2 and Fig. 3B), different to primitive and extraembryonic endodermal tissues[7, 16]. One of these key marker is *SOX17*, a TF of the SOX superfamily has been implicated in diverse molecular and developmental processes of several developmental model systems[10, 30-35]. Interestingly, so far *SOX17*'s role in the formation of human *in vitro* definitive endoderm is unclear and its function and relevance for human gastrulation remains understudied.

### 1.1.3 Phylogeny and function of SOX genes

*SRY*-related high-mobility-group (HMG)-box (SOX) genes encode for a TF superfamily binding to the minor groove in DNA. Their main-characteristic is a homologous and highly conserved sequence, called the HMG-box. Members of the SOX superfamily are found across the animal kingdom and are involved in a range of very diverse developmental processes.[36]

1990, the SOX family – which nowadays is composed of 10 different groups (A-J) plus outgroups – was originally identified based on the conservation of the HMG-box for the sex determining region of Y-gen (*SRY*), the mammalian testis-determining factor [37]. Per

definition, SOX proteins as such have to be more than 50% identical to SRY in the HMG domain. According to Bowels J. et al., that definition seems to be incorrect in regards of newly identified SOX genes which do not follow this rule. The classification based on a strict 50% identity to SRY-HMG is a historical, arbitrary, and, in retrospect, poor choice for such SOX family comparisons. SRY has arisen only in the mammalian lineage and is clearly very divergent, hence the SRY-HMG domain comparison is not a suitable indicator of SOX family membership. In comparison, vertebrate orthologues are highly conserved in HMG domain sequence. Outside of the HMG sequence, this high conservation falls off considerably [38-42]. The study by Bowels J. et al. provides an alternative criterion to define SOX genes using the conservation of key motifs within the HMG domain. This sequence, RPMNAF at position 5–10 is conserved for all SOX sequences, including those of groups H, I, and J, but not for the most closely related outgroups fu-MATA1, mo-LEF1, and mo-TCF1. Interestingly, the motif is also present in one of the SOX-like genes in *Drosophila melanogaster*, *capicua* (*cic*), which has apparent orthologues in *Caenorhabditis elegans* and humans, suggesting that this 6-amino-acid motif is insufficient to strictly define SOX genes[43]. The extended version however is common to all non-SRY SOX members (RPMNAFMVW) and appears to be the most reliable signature of the SOX family.[36]

When it comes to evolution of the SOX family, there is evidence by both slow divergence and the recruitment of preexisting functional elements. The HMG domain, the ancestral motif which forms the core of SOX family proteins, is expected to have gradually accumulated sequence changes under the selection pressure of retaining sequence-specific DNA-binding function. In contrast, variability of SOX proteins outside of the HMG domain indicates that sudden and stochastic evolutionary changes must also have occurred apparently via co-option of functional domains and motifs resulting in the formation of “evolutionary chimeras”[44-47]. Such changes may, at least to some extent, mark the origin of the various SOX groups. Subsequent to these major changes, additional duplication and divergence events must have occurred, resulting in the range of SOX proteins present in vertebrates today.[36] Throughout the mouse and human genomes, SOX genes are widely dispersed [48], arguing against a purely tandem duplication model of SOX family expansion. It is proposed that the family has arisen from a common ancestor via ancient duplication, dispersal, mutation, and acquisition mechanisms. Hence, throughout metazoan evolution, HMG box-containing sequences may have been duplicated according to the duplication– degeneration– complementation model[49, 50].

The idea is that throughout evolution SOX genes were duplicated, in each case leaving one redundant copy which was then free to evolve a new function or else be lost from the genome, also called as subfunctionalization[49, 51]. Subfunctionalization can occur in two different ways. Following duplication and individual degeneration of daughter copies, both daughter copies will acquire joint evolutionary beneficial functions (“joint subfunctionalization”). The

alternative is that a temporal expression pattern of the ancestral gene is maintained by the duplicate daughter genes (“temporal subfunctionalization”). For SOX genes both might be holding through noting their close phylogenetic relation among genes within a certain group[36] and their tissue specific expression patterns in early development, as for instance in definitive endoderm[17]. To facilitate tissue specificity and tightly controlled spatiotemporal expression of SOX factors a variety of sophisticated gene regulatory mechanisms have evolved, which guarantee proper gene control and ultimately correct development of multicellular organisms.

## 1.2 Gene regulation in development

All species across the animal kingdom share a basic set of genes which regulate and orchestrate their development. To do so, spatiotemporal precision is highly critical for proper embryogenesis and the establishment of their body plan. To become a multicellular organism a variety of mechanisms evolved, on the cellular, molecular, and genetic level. These mechanisms are key and a consequence to gain a highly sophisticated transcriptional apparatus.

The initial sequencing and analysis of the human genome[52] revealed that only 2-3% of the mammalian genome contains protein-coding regions, with the remaining 97% non-coding DNA (ncDNA) have no immediately clear function[53]. Nevertheless, recent investigations have shown that ncDNA or so called “junk” DNA harbors *cis*-regulatory elements (CREs) which are critical for gene regulation and control[54, 55]. Along these lines it is interesting to note that the number of genes within different lineages has remained comparatively stable while CREs have expanded, diversified, and altered during millions of years of evolution. Therefore, CREs are important for the development of complex gene regulation patterns with changes in their activity driving altered target expression and evolutionary novelty. During the last decades many insights into how CREs facilitate complex gene regulation has emerged. Nevertheless, the in-depth mechanisms of how their communication works is poorly understood. To understand how and which factors determine the emergence of gene expression patterns, it is inevitable to study single loci gene regulation in well-established model systems as the *SOX17* locus in the context of *in vitro* derived definitive endoderm, as utilized in the studies of this thesis.

### 1.2.1 The non-coding genome: *cis*-regulatory elements (CREs)

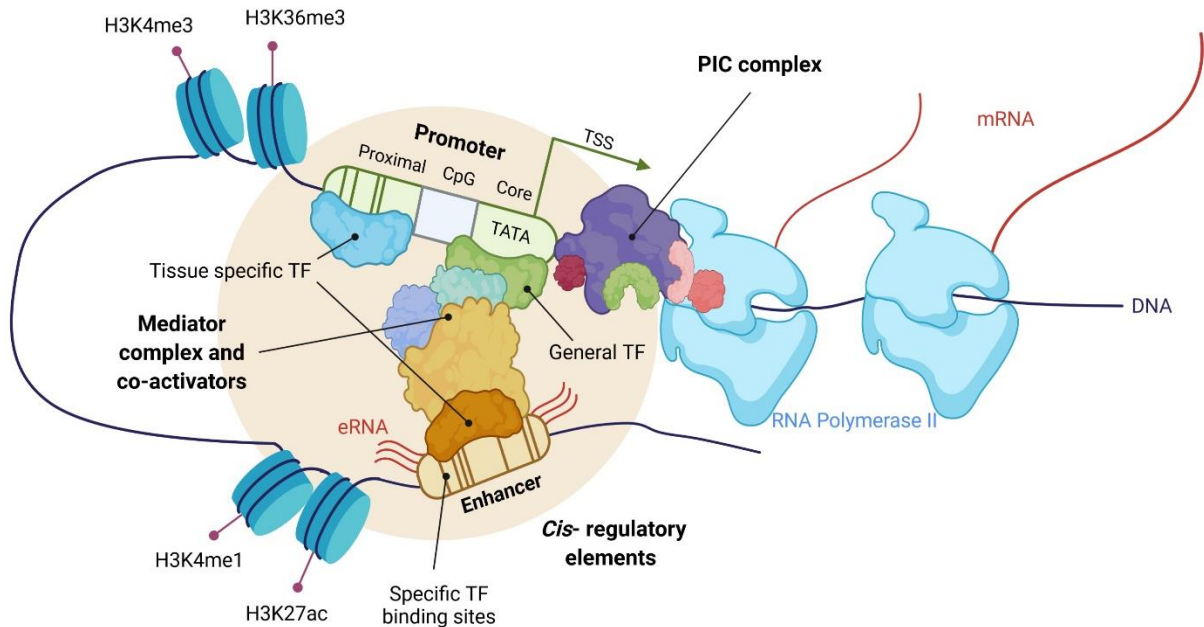
The non-coding genome performs many essential functions, especially in the regulation of gene expression, but it is unclear yet how much of it is necessary. Functional components of ncDNA regulatory sequences are called CREs, including enhancers, core promoters, matrix or scaffold attachment regions, insulators and silencers[56]. Historically, CREs have been generally placed into two classes, promoters and enhancers, that each distinctly drive gene expression in an orchestrated organization. However, CREs can also be repressive elements like silencers and insulators. For example, insulators are DNA elements binding insulating proteins to block the effect of an enhancer when positioned between an enhancer and its target promoter[57, 58]. The best characterized insulator-binding protein is the 11-zinc finger protein CCCTC-binding factor (CTCF), which has been associated with repressive activity. However, the biochemical mechanisms of this repressive activity are only recently beginning to be understood[59, 60]. Unfortunately, there is no general definition of a CRE. The best working definition follows one given to enhancer[61]: a *cis*-regulatory element such as an enhancer is

defined as the smallest fragment of DNA that, when linked to a reporter gene and transferred into an appropriate cell, executes a regulatory function in a fashion consistent with that of the native gene in its proper context. This definition evidently simplifies natural relations between CRE structure and its function by paring down *cis*-neighboring sequences. It is not possible to define precise borders (up to one nucleotide) for any class of CREs. Besides the increasing research interest into repressive CREs, the studies of this thesis will mainly focus on the most studied CRE classes, namely the activating elements promoters and enhancers.

### 1.2.2 Promoters and Enhancers

CREs, in particular activating promoter and enhancer elements as the most studied elements among them, regulate spatiotemporal gene expression as noted earlier. Varying in size from few to thousands of base pairs (bp), both are defined as DNA elements that can activate transcription. Dispersed throughout genomes, they are both highly conserved across species and one of their most common features is their high abundance of transcription factor (TF) binding sites whose facilitate transcription (Fig. 4). Either way, both types of elements generally display some unique characteristics.

As for example, mammalian promoters are composed by core and proximal components (Fig. 4). Core promoters which are comprised by general TF-binding sites, support the formation of the preinitiation complex (PIC) of RNA-Polymerase II (PolII) in proximity of the transcriptional start site (TSS) within 50-100 bps[62, 63]. In general they are known to be basally active, they bear common motifs as the TATA-box and very often contain CpG-islands (70%)[64, 65]. Importantly, their activity or repression potential is mostly governed by CREs. Proximal promoters in contrast are located more far from the TSS and serve as a platform for binding of tissue-specific TFs. So far, it is unclear how the proximal promoter communicates with the core in a functional manner and how it confers to gene activation. One proposed mechanism may be communication with activating regulatory sequences like enhancers to ultimately release assembled RNA PolII complexes to initiate active transcription[62]. Until today there is no conclusive characterization of proximal promoters, although the depiction as core promoter-neighboring enhancers seems reasonable (Fig. 4). Functionally, promoters mainly act in a unidirectional but are also reported to act in a bidirectional way. Bidirectional promoters[66] have been shown to often act as strong enhancers too, while unidirectional promoters generally cannot. The balance between enhancer and promoter activity is generally reflected in the levels and directionality of enhancer RNA (eRNA) transcription and is likely an inherent sequence property of the elements themselves[67] (Fig. 4).



**Fig. 4 Transcriptional apparatus governed by promoter and enhancer elements.** Scheme of a transcriptionally active region whose enhancer is occupied by tissue specific TFs (orange) and marked by active enhancer histone modifications (H3K4me1, H3K27ac). Note how the distal enhancer DNA element “loops” near the promoter-core of the coding gene via Mediator complex and co-activators that are bound to general TFs (green). Also note the strong enhancer is lowly transcriptionally active, producing undirected eRNAs. Transcription is shown to be finally initiated via the assembly of the Polymerase II (PolII) pre-initiation complex (PIC) followed by RNA PolII recruitment to the promoter of the coding gene and mRNA-transcription of the coding DNA-sequence. The transcriptionally active promoter and the gene-body are marked by histone modifications highlighting active ongoing transcription and gene-body identity respectively (H3K4me3, H3K36me3). (Created with BioRender.com)

Different to promoters, enhancers locate highly variable. The most famous transcription-enhancing DNA sequence was discovered in 1981 [68, 69]. This element – a 72 bp sequence from the SV40 virus – was identified to significantly upregulate transcription of the rabbit beta-globin gene *in vitro* episomally. That this type of DNA sequence would have cell-type and developmental-stage specific activities was found later and only then was termed an enhancer. The best working definition for enhancers, besides their varying positioning and the absence of a TSS (Fig. 4), is the smallest fragment of DNA that, when linked to a reporter gene and transferred into an appropriate cell, executes a regulatory function consistent with that of the native gene in its proper context. Naively, enhancers were not expected to have the necessary sequences for transcription initiation themselves. Nevertheless, well-characterized enhancers have been observed to be transcribed [70]. This is in line with the finding of PolII and general TFs binding ability, suggesting some inherent promoter activity [71].

Promoters and enhancers are crucial for the regulation of genes which can be mapped and validated by various techniques. However, how an enhancer induces transcription remains largely unclear. As enhancers are thought to function in a position- and orientation-independent manner to activate their target genes [72] and can be located very distal to their cognate promoter [72], physical proximity is required to initiate transcription.



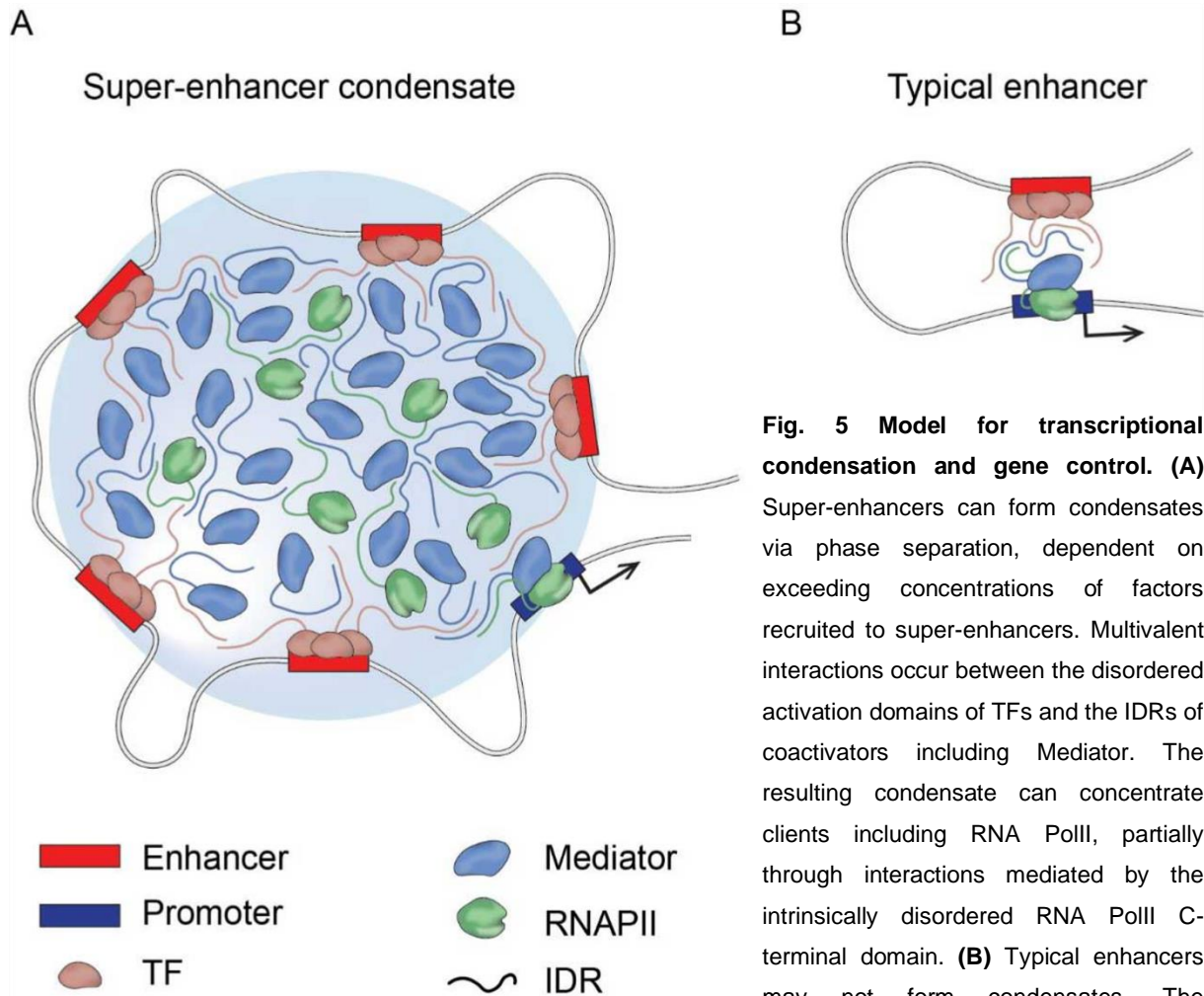
### 1.2.3 Regulation by distal enhancers

The regulation of genes by enhancers via long distances occurs mostly for developmental genes, with linear distance ranges of 1-1.145 Mb[73, 74]. Commonly its thought that distal enhancer-promoter interactions are facilitated via DNA-looping in the nuclear space to activate gene expression[75]. These chromatin loops facilitating long-range interactions are mostly observed in regulatory gene landscapes. So far there are several different models of loop-facilitation further leading to gene-activation that were proposed[76]. The most recent model hypothesized for enhancer function is based on homotypic interactions between tissue-specific TFs bound to both elements, to bridge their interaction and recruit the transcriptional apparatus (Fig. 4).

Enhancers contain specific TF binding motifs, like proximal promoters (Fig. 4 and Fig. 5B). These factors facilitate active transcription through the integration of complex information of their target genes via transacting functions[77, 78]. TFs, whose recognize and bind individual DNA motifs of promoters and enhancers tissue-specifically recruit co-activators including Mediator and P300, necessary for the further recruitment of the transcriptional apparatus (Fig. 4 and Fig. 5B)[72, 79]. Interestingly, recent investigations have shown that liquid-liquid phase separation may be a possible mechanism for TF-mediated gene activation besides enhancers just serving as being binding platforms. In this particular circumstance transcription is suggested to be induced by the formation of active membrane-less hubs incorporating high concentrations of the transcription apparatus, containing TFs, Mediator and RNA PolII especially at clusters of enhancers, so called super-enhancers (SEs)[80-82] (Fig. 5A). The formation of those membrane-less condensates have been described to cause gene-activation for pluripotency factors as e.g. OCT4 in human iPSCs[83]. Interestingly, altered capacity of HoxD13 intrinsically disordered region (IDR) for example impairs transcriptional regulation *in vivo* and has been shown to cause congenital malformation[84].

In very recent studies, it is indicated that transcription is a non-equilibrium process that provides dynamic feedback through its RNA product. Surprisingly, these results hind to an extended model where RNA is providing a positive and negative feedback on the actual act of transcription through the regulation of electrostatic interactions in transcriptional condensates. The formation of transcriptional condensates includes accumulation of TFs by enhancer DNA[85] and electrostatic and other interactions between the IDRs of TFs and coactivators (Fig. 5A)[82, 83], which seems to drive RNA promoting and dissolving condensates[86]. The model proposes that, lower levels of short RNAs produced during transcription initiation

promote formation of condensates, while higher levels of longer RNAs during elongation can cause condensate dissolution.



**Fig. 5 Model for transcriptional condensation and gene control. (A)** Super-enhancers can form condensates via phase separation, dependent on exceeding concentrations of factors recruited to super-enhancers. Multivalent interactions occur between the disordered activation domains of TFs and the IDRs of coactivators including Mediator. The resulting condensate can concentrate clients including RNA PolII, partially through interactions mediated by the intrinsically disordered RNA PolII C-terminal domain. **(B)** Typical enhancers may not form condensates. The concentration of factors recruited to typical enhancers may not reach the threshold required for phase separation. (Adapted and modified figure from Sabari R.B. et al., 2018)

Taken together, CREs and their associated TFs are crucial for spatiotemporal regulation of precise gene expression for cell fate decisions and tissue specification during development of multicellular organisms. How long-range interactions are established, and an enhancer-element can find and regulate its target promoter remains elusive. Hence, to understand long range linear distance communication between elements and to understand which enhancer may regulate which gene in a certain context seems challenging and key investigation for the future. Recent studies shade light on RNAs and the act of transcription itself which may play a key role in long-range interactions in transcriptional condensates.

### 1.3 The 3-dimensional (3D) genome

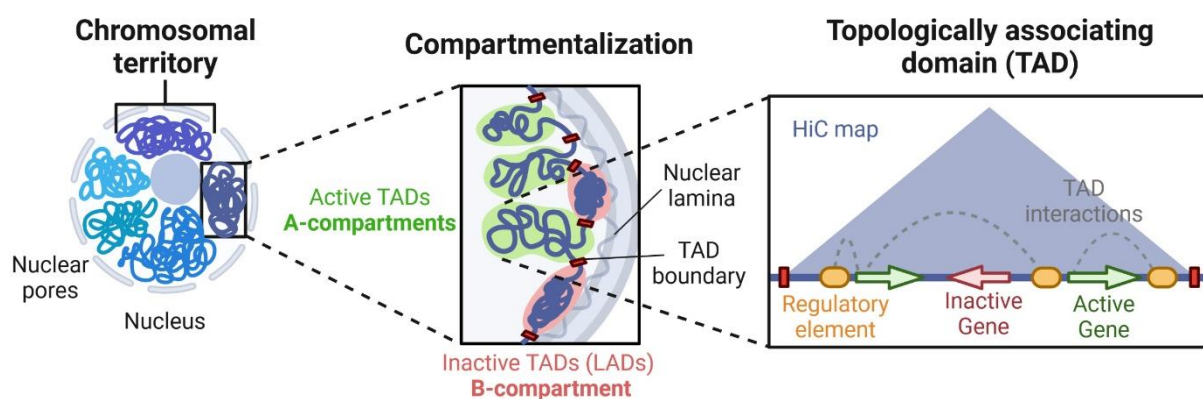
Several studies highlighted the importance for physical proximity of enhancers and their target specific promoters for gene control and activation[87-91]. Nevertheless, it remains unclear how this proximity is controlled mechanistically. More and more studies over the past decades revealed insights into the connection of 3D organization of chromatin with the transcriptional regulation of target genes, linked to the functional assignment by CREs. To study genome organization, there is a complementary variety of methods existing to provide distinct insights into 3D genome architecture. Various bulk-techniques as e.g., proximity ligation assays addition single-cell based methods as e.g. DNA fluorescence *in situ* hybridization (DNA-FISH)[92]. As it became evident from 1960 onwards that DNA within the nucleus is highly organized into hetero- and euchromatin, further investigations identified these compartments associated with the nuclear periphery and the center of the nucleus respectively. These studies were fundamental and are the basis for nowadays understanding of nuclear spatial organization. The identification of chromosome territories revealed a non-random fashion localization within nuclear space by chromosomes[93, 94]. Ever since, the impact of spatial organization and DNA positioning on gene regulation was kept in focus[95, 96]. Advances in DNA sequencing technology, especially the invention of Chromosome Conformation Capture (3C)-based methods, significantly increased the understanding of 3D chromatin organization on a locus-level and describe the physical association of enhancers and their target promoters[97]. In particular, to identify DNA-DNA and DNA-protein contacts chromatin conformation capture (3C) -methods utilize reversible crosslinking followed by digestions and proximity ligation to test for closer proximity of two genomic loci. The formed circular hybrid molecules whose frequency of abundance stands for interaction frequency, is further quantified by quantitative real-time polymerase chain-reaction (qRT-PCR) (one vs. one, 3C). Utilizing this technique, genome-wide interaction frequency maps can be generated. So far proximity ligation has been serving as the principle for the generation and refinement of various technologies[98]. Combining 3C with high-throughput sequencing techniques led to the invention of proximity ligation-based methods further developed and refined as e.g., circular chromatin conformation capture (one vs. all, 4C). 4C sequencing, which is based on an additional PCR-enrichment step after proximity ligation followed by sequencing, generates an interaction frequency map of the genome for a specific genomic location of interest (viewpoint). Including an additional RNA-probe enrichment step targeting many viewpoints of a library followed by sequencing is utilized by Capture-C sequencing (many vs. all). Finally, sequencing of all obtained ligation fragments generates a genome-wide interaction map (all vs. all, Hi-C). Capture Hi-C (cHi-C) is a combined derivative of Hi-C and Capture-C which creates high-resolution interaction frequency maps of an enriched genomic region of interest (up to 5 Mb) without sequencing all obtained fragments.

### 1.3.1 Chromosome compartments

Over the last decades it has been shown that chromosomes exhibit numerous inter- and intrachromosomal organizational layers of interphase chromosomes (Fig. 6 left). This type of organization has mainly been identified by studies utilizing genome-wide Hi-C sequencing. Interestingly, it was observed compartmentalization into active (A-compartments) and repressed (B-compartments) compartments (Fig. 6 center), which are higher-order chromatin segregates. Compartmentalization has been found to highly dependent on criteria as e.g., replication-timing, compaction (Eu- vs. Heterochromatin), general transcriptional activity, the association with the nuclear envelope (Lamina-associated domains, LADs) and their epigenic state. Hence, active and inactive compartments form homotypically according to those criteria[99].

### 1.3.2 Topologically associating domains (TADs)

Follow up investigations elucidated partitioning of chromosomes along their length into so called topologically associating domains (TADs), which are sub mega base units [100, 101] (Fig. 6 right). These architectural units, surprisingly overlap largely with regulatory landscapes of developmental genes, shown by sensor insertions[102, 103]. These studies indicate that gene-specific regulatory information is retained until sensors get inserted into a demarcated topologically associated neighborhood. Hence, the association of enhancers and promoters is suggested to be assisted by TADs. TADs size average around 850 kb, are highly conserved across species and generally defined by their preferential intra-TAD interactions[100, 101].



**Fig. 6 Scheme of the 3D-genome organizational layers.** From left to right depicted are chromosomal territories, the compartmentalization of chromosomes, and their topologically associated domains (TADs). Chromosomes within the nucleus locate at distinct territories (left). A- and B- compartments are further segregated into active and repressed regions on chromosomes (center). The genome is further partitioned into TADs (green) and LADs (red). Boundary separated TADs are further displayed in detail (right). Active genes (green) and non-active genes (red) within the TAD are regulated by their regulatory elements (CREs, orange). Intra-TAD interactions are indicated by dashed lines. (Created with BioRender.com)

However, TADs provide only limited insight into the molecular mechanisms causing specific enhancer-gene interactions, containing on average about 8 genes whose expression is weakly

correlated[104]. A collection of data provided by several studies suggested TAD sub-domains, mainly CCCTC-binding factor (CTCF)/Cohesin anchored[91, 105-107]. These so called CTCF loop domains or insulated neighbourhoods[104, 107], with a median size of 190 kb contain 3 genes on average. Interestingly, these domains were unveiled to carry very often (~40%) only a single gene (TIG). They were also found to mainly enrich for developmental regulators with highly correlated enhancer activity and strongly conserved CTCF boundaries across mammals and different human tissues[91].

Nevertheless, TADs are also insulated by each other through boundaries enriching for DNA-bound CTCF and structural maintenance of chromosomes (SMC) complex [100, 101, 108-113]. But the role of CTCF in TAD maintenance and establishment remained unclear for some time and it is still unclear what TADs physically represent. The most accepted model which is strongly supported by computational polymer-modelling, is the ATP-dependent loop extrusion mechanism[114, 115], where the formation of loops is highly dependent on towards oriented CTCF-motifs at TAD-boundaries facing each other[116-118]. Cohesin, constantly extrudes the chromatin fiber bidirectionally after Nipbl-dependent loading onto the DNA. Cohesin is thought to bypass CTCF motifs oriented in active extrusion direction until stopped by CTCF-motif bound towards[114, 115]. Cohesin, also dynamically dissociates from chromatin, mediated by the releasing factor Wapl which highlights the constant turnover of structural proteins bound to the chromatin fiber[119-121].

Moreover, LOF studies of either CTCF or Cohesin, demonstrated a genome-wide TAD-loss *in vitro*, indicating their relevance for TAD formation[122-124]. Taken together, due to the high correlation with gene regulatory landscapes, TADs are believed to provide an insulated microenvironment for enhancer-promoter communication, shaped by CTCF and Cohesin, whose functional relation between remains still elusive.

### 1.3.3 Intra-TAD interaction dynamics

TADs which are defined by Hi-C represent domains that are insulated and according to sensor studies seem to function as regulatory units. Nevertheless, how the actual structure within TADs may facilitate the communication between enhancers and promoters is barely understood. So far there are two models of three-dimensional controlled gene expression proposed, the permissive and instructive model. The permissive model proposes a state in which enhancers and promoters within TADs stay in an unaltered state of proximity by the 3D architecture until activated by tissue-specific TFs. Whereas the instructive model describes tissue specific *de novo* interactions or chromatin-looping followed by immediate transcriptional activation of a gene[125]. The direct impact of CTCF in both remains controversial. Interestingly two types of contact-modes have been observed which are matching above models. Interactions within TADs indeed seem highly dynamic throughout differentiation and

development within a certain cell population shown by single cell studying techniques as e.g. microscopy and other single cell studies compared to Hi-C, which is a bulk-method[126-128]. These tissue specific dynamic intra-TAD interactions and TAD formations have been shown in high-resolution Hi-C of differentiated mESCs to NPCs[129] and similar results have been obtained in immune cell trans-differentiations[130]. On the other hand, matching the permissive model numerous genome wide enhancer-promoter loops at various loci have been revealed by high-throughput techniques as e.g., Hi-C, micro-C and Hi-chromatin immunoprecipitation (Hi-ChIP)[131-133]. Showing that tissue-invariant structures may be crucial to sustain regulatory function in some cases, it has been shown that disruption of CTCF-sites at the edges of the Shh TAD, where Shh and its limb enhancer – zone of polarizing activity regulatory sequence (ZRS) – are maintained in close proximity leads to Shh expression reduction in limb bud development[89]. Taking together these results indicate that in order to sustain full regulatory function tissue-invariant structures may be important.

#### 1.3.4 The role of CTCF in 3D chromatin organization

CTCF and Cohesin which are mostly found at TAD-boundaries are also present within TAD substructures forming invariant loops also referred as loop domains[100, 109, 112, 129]. These are thought to support enhancer-promoter interactions by sampling of the regulatory environment at promoters for a suitable enhancer. CTCF-dependent tissue-specific interactions, however have rather been described for long-range than *de novo* contacts in invariant loop-domains[112]. These dynamic enhancer-promoter contacts seemingly do not rely on CTCF during differentiation, rather establishing tissue-specific TF and/or polycomb repressive complex 1/2 (PRC1/2) interaction networks[129, 134]. These results have recently been confirmed in neural differentiation where CTCF-dependent enhancer-promoter contacts promote long-range interactions. Here CTCF sites proximal to a promoter seemingly serve as anchors to reel in potential regulatory elements for gene activation[135]. Moreover, recent investigations have identified two distinct regimes of RNA-dependent and RNA-independent CTCF-anchored loops. Here, deletion of the RNA-binding domain (RBD) of CTCF led to genome wide CTCF decrease in mESCs, which potentially explains tissue-specific domain formation during development and stabilization of long-range contacts[136]. Not only active transcription but also DNA-methylation is a regulatory mechanism, crucial for tissue-specific CTCF-dependent loop formation as shown for IDH mutant gliomas, where reduced CTCF-binding led to loss of insulation between TADs and ectopic oncogene activation finally resulting in tumorigenesis[137]. Prior to that knowledge, CTCF has been demonstrated to form various invariant and dynamic chromatin loops matching either of the above models[125]. Nevertheless, it is not fully clear how CTCF shapes the regulatory landscape of developmental genes and impacts gene-control.

## 1.4 Non-coding RNAs (ncRNAs)

As outlined in the previous chapters, CREs and their associated TFs are crucial for spatiotemporal regulation of precise gene expression for cell fate decisions and tissue specification during development of multicellular organisms. How long-range interactions are established, and an enhancer-element can find and regulate its target promoter remains elusive. Hence, to understand long range linear distance communication between elements and to understand which enhancer may regulate which gene in a certain context seems challenging and key investigation for the future.

Recent studies shade light on RNAs and the act of transcription itself which may play a key role in long-range interactions in transcriptional condensates. In the context of phase-separated condensates, transcription is a non-equilibrium process that provides dynamic feedback through its RNA product. Interestingly, these results hint to an extended regulatory model where RNA is providing a positive and negative feedback on the actual act of transcription through the regulation of electrostatic interactions in transcriptional condensates. The formation of transcriptional condensates includes accumulation of TFs by enhancer DNA[85] and electrostatic and other interactions between the IDRs of TFs and coactivators[82, 83], which seems to drive RNA promoting and dissolving condensates[86]. The model proposes that, lower levels of short RNAs produced during transcription initiation promote formation of condensates, while higher levels of longer RNAs during elongation can cause condensate dissolution. Moreover, there are also two distinct regimes of RNA-dependent and RNA-independent CTCF-anchored loops which so far could be identified. Interestingly, deletion of the RNA-binding domain (RBD) of CTCF led to genome wide CTCF decrease in mESCs, which potentially explains tissue-specific loop domain formation during development and stabilization of long-range contacts[136]. Therefore, RNA and the act of active transcription clearly play crucial roles in participating into gene-regulation and 3D chromatin architecture, being of interest for future investigations.

Numerous studies have demonstrated that the true catalog of RNAs encoded within the mammalian genome (the “transcriptome”) is more extensive and complex than previously thought[138-140]. In humans and mice, for instance, it has become apparent that the vast majority of the genome is transcribed, often in intricate networks of overlapping sense and antisense transcripts, many of which are alternatively spliced[138, 141-144]. However, mRNAs account for only ~2.3% of the human genome[138, 145], and therefore the vast majority of this unexpected transcription, sometimes referred to as “dark matter”[146, 147], appears to be non-protein coding or so called non-coding RNAs (ncRNAs). Among the broad variety of ncRNAs, enhancer RNAs (eRNAs) and long non-coding RNAs (lncRNAs) have frequently been implicated with transcriptional control and gene-regulation[148, 149].

### 1.4.1 Enhancer RNAs (eRNAs)

Enhancer RNAs (eRNAs) which are product of enhancers active transcription, were very well studied in mouse and human cells and were found to share common properties but which role they may play in organizing 3D genome architecture and the formation of condensates is not well understood. In mammalian, transcribed eRNAs are lowly abundant, non-polyadenylated, unspliced and retained in the nucleus[150-152]. eRNA loci are mainly transcribed bidirectionally[153, 154] (Fig. 4) although some have been reported to be transcribed unidirectionally as well[71, 150]. The underlying mechanisms of how eRNAs facilitate enhancer function remain uncertain. In many instances it may be that the eRNA has no function and is a consequence of RNA PolII binding to open chromatin rather than acting by its own. As a possibility with ncRNAs, it also may be that the act of transcription of the enhancer per se, and not the RNA itself, is important for enhancer function. This has been shown earlier, where the actual transcript is not relevant as it is equally capable of maintaining transcriptional activation in either the sense or antisense orientation[155, 156].

Interestingly, mammalian genomes are comprised of loci generating a great diversity of noncoding RNA classes, but the defining criteria for each class are not always obvious. Previous investigations[157] have challenged the distinction between lncRNAs and eRNAs in mouse erythroblasts, and provides an experimental approach to define their mechanism of action. In this study the authors show that a lncRNA *Lockd* (long, stable, commonly spliced and polyadenylated, and transcribed from its own promoter) is an actual eRNA, whose CRE serves as a proximal enhancer for the close by gene *Cdkn1b*. By insertion of early polyadenylation signals into the *Lockd* 5' region downstream of its promoter (*Cdkn1b* CRE), the authors show that *Cdkn1b* expression is unaltered, and the *Cdkn1b* promoter retains physically interactions in *cis* with the *Lockd* CRE. Perturbations of the *Lockd* CRE instead affect the expression of *Cdkn1b* in fact highlighting the act of active transcription of the CRE regulating *Cdkn1b* gene control. This case study not only provides a great experimental framework to disentangle different novel RNA species but also an important perspective on the evolution of eRNAs and lncRNAs, in which eRNAs may beget lncRNAs with *trans* function.

### 1.4.2 Long non-coding RNAs (lncRNAs)

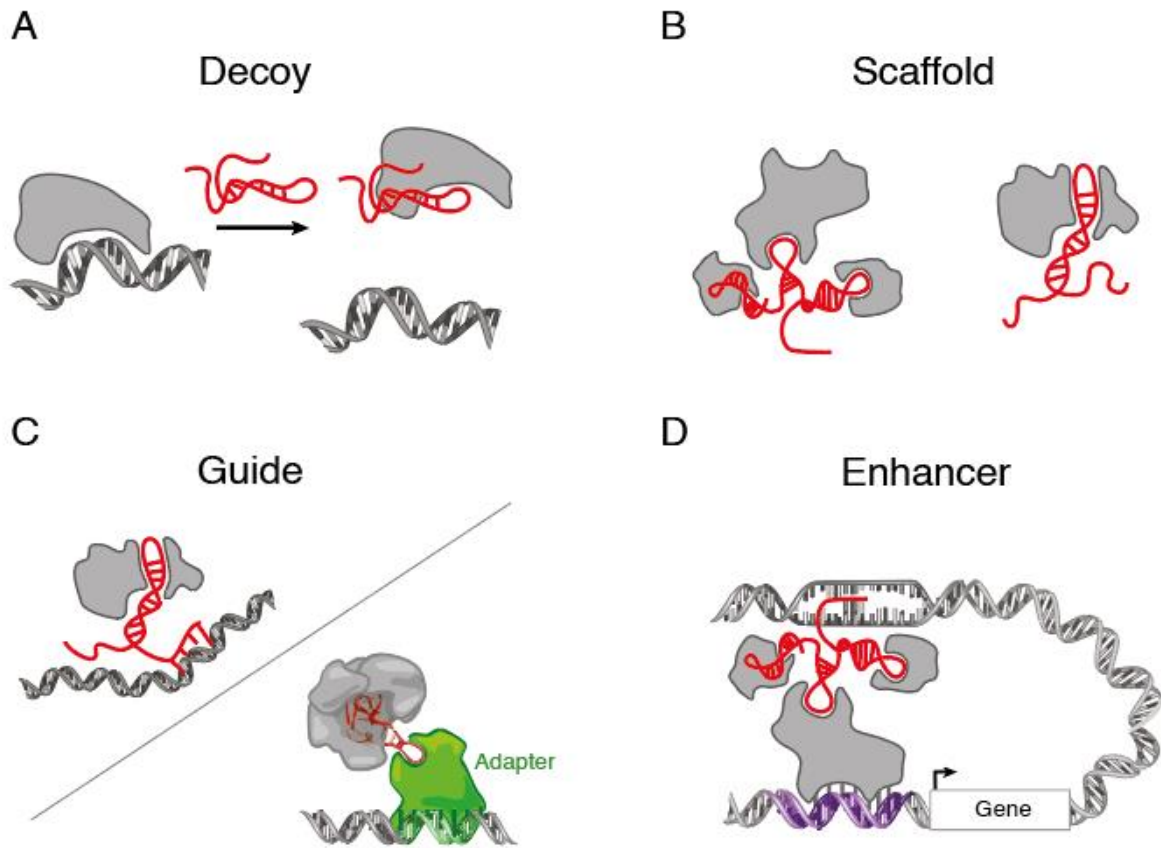
lncRNAs are broadly defined as RNA transcripts  $\geq 200$  nucleotides (nt), which are not translated into functional proteins, with recently described exceptions of micropeptides[158] or microproteins[159]. This broad definition includes a big number of a highly heterogenous group of transcripts, which differ in their biogenesis and genomic origin. Initially, it has been suggested that the human genome encodes for more than 16,000 lncRNAs[160] while other estimates exceed 100,000 human lncRNAs[161, 162]. Included among these are lncRNAs, mainly transcribed by PolII but also other RNA polymerases and lncRNAs from intergenic



regions (lincRNAs) as well as sense or antisense transcripts overlapping with coding genes. PolIII driven lncRNAs share very similar biogenesis as mRNAs, as they are often capped by 7-methyl guanosine ( $m^7G$ ) at their 5' ends, bear polyadenylated 3' ends and they are spliced by the same machinery. However, they also differ in many characteristics compared with mRNAs. LncRNAs have generally lower expression levels, less sequence conservation, are localized mainly in the nucleus, and splicing appears to be less efficient[163, 164]. One of the most surprising findings is that lncRNAs can exhibit very little sequence conservation yet retain critical evolutionary conserved functions.

Recent discoveries showed that lncRNAs are important components of the regulatory network in the genome[165] but how lncRNAs may exert this and other functions still remains elusive. Hence, the question how lncRNAs function remains. Several hypotheses how lncRNAs selectively interact with the genome have been proposed, including formation of an RNA-DNA-DNA triplex, RNA binding to a sequence-specific DNA-binding protein, a RNA-DNA hybrid that displaces a single strand of DNA (so-called R-loop) and an RNA-RNA hybrid of lncRNA with a nascent transcript[166, 167].

The interaction and binding of proteins equips lncRNAs with several regulatory capacities. Despite our limited knowledge from just dozens of characterized examples, several mechanistic themes of lncRNAs' functions have emerged (Fig. 7)[168]. Decoys: First, and at the simplest level, lncRNAs can serve as decoys that preclude the access of regulatory proteins to DNA[169, 170]. Scaffold: The lncRNAs can serve as adaptors to bring two or more proteins into discrete complexes[171]. Guides: Many lncRNAs are individually required for the proper localization of specific protein complexes[172-174]. Enhancer: Bridging distal DNA-elements such as promoters and enhancers via protein-complexes. Besides eRNAs, previous studies performed loss-of-function (LOF) experiments and found 7 of 12 lncRNAs affected expression of their cognate neighboring genes[175]. The authors continued to demonstrate it was not the act of transcription rather the RNA itself that was important for gene enhancer activation. Although this trend of lncRNAs affecting transcription of neighboring genes is not a universal phenomenon[169, 176], these studies clearly demonstrate a functional role for the RNA molecule beyond that of a simple by-product of transcription in enhancer regions. Despite the modes of action described, the question remains how certain lncRNAs may function in the biological context of development and yet, their biological relevance needs to be further investigated.



**Fig. 7 Modes of lncRNA action.** (A) The lncRNAs can act as decoys that titrate away DNA-binding proteins, such as transcription factors. (B) These lncRNAs may act as scaffolds to bring two or more proteins into a complex or spatial proximity and (C) may also act as guides to recruit proteins, such as chromatin modification enzymes, to DNA; this may occur through RNA-DNA interactions or through RNA interaction with a DNA-binding protein. (D) Such lncRNA guidance can also be exerted through chromosome looping in an enhancer-like model, where looping defines the *cis* nature and spread of the lncRNA effect. (Adapted and modified figure from John L. Rinn and Howard Y. Chang, 2012.)

## 1.5 The human *SOX17* locus in definitive endoderm

Various developmental studies utilize *SOX17* as a key-marker for the identification of human definitive endoderm[21, 177-179] but no one has studied the effects of loss of *SOX17* in early human definitive endoderm, hence its role during the formation of that germ-layer due to loss of function still remains elusive. Also, little is known about the regulation of the *SOX17* locus during the formation of human definitive endoderm.

Engert S. et al., has investigated the mouse *Sox17* proximal upstream and downstream regulatory regions and identified Tcf/Lef binding elements (TBEs). In this study it was found 13 TBEs in the 8945 kb upstream and downstream regulatory region, including intron 1 and 2. In mouse endoderm *Sox17* has several alternative transcription start sites that are used in a vascular endothelial and endoderm tissue-specific manner[180-182]. Both Tcf4 and  $\beta$ -Catenin were occupied in ESCs that had been induced by Wnt3a and activin to differentiate into endoderm but were not bound to the TBEs under pluripotency conditions which suggested that canonical Wnt signalling regulates *Sox17* via Tcf4/ $\beta$ -catenin complexes during endoderm formation. Recent investigations in the context of ovarian cancer by Reddy J. et al. have highlighted a H3K27ac decorated distal region upstream of the *SOX17* locus, co-occupied by cardinal tumor-suppressive TFs as PAX8/MECOM and *SOX17* itself[183]. Their data suggested a distal SE upstream of *SOX17*. These findings are highly in concordance with the earlier reported definitive endoderm specific differentially methylated region (DMR) 230 kb upstream of *SOX17* by Tsankov A. et al., overlapping with the SE[29].

### 1.5.1 Definitive endoderm and its DNA methylome

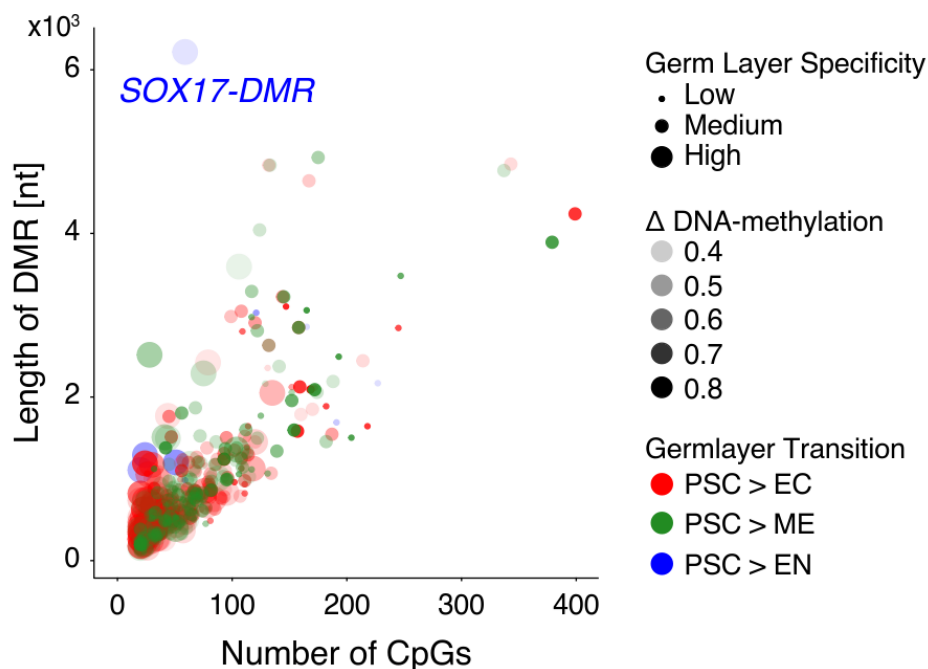
Interestingly, when comparing the DNA methylomes from undifferentiated and terminally differentiated cell states of the *in vitro* germ-layer, a lot of epigenetic changes occur throughout their genomes[21, 29], one of them being loss of 5-methylcytosine (5mC) methylation levels the at CpG dinucleotide motif, genome wide (Fig. 8).

DNA methylation which is the biological process of methyl groups being added to DNA (CpG 5mC)[184], can change the activity of a DNA segment without changing the actual sequence[185]. In mammals, around 75% of CpG dinucleotides are methylated in somatic cells[186]. Despite recent findings of the DNA methyltransferase enzyme DNMT1 having *de novo* activity[187], DNMT1 catalyzes maintenance of 5mC of hemimethylated DNA during replication[188]. Besides that, in early embryonic development, establishment of the erased methylome is mainly catalyzed by the *de novo* enzymes DNMT3A and DNMT3B[189]. The DNA methylation landscape of mammal somatic tissue is very particular as it appears in a global bi-modal fashion of overall high CpG methylation and low at gene-bodies and promoters. When located in gene promoters, DNA methylation typically acts to repress gene transcription[190]. In mammals, DNA methylation is essential for normal development[191] and

is associated with several key processes including genomic imprinting, X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis.

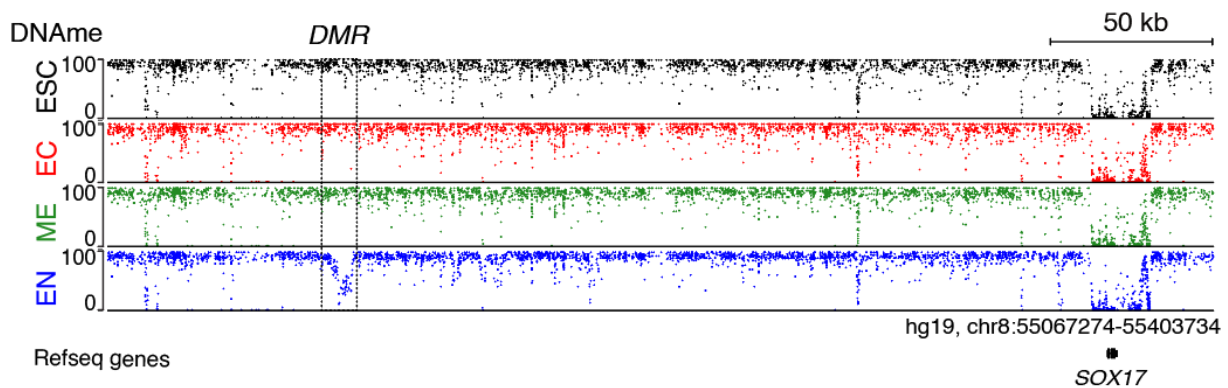
When comparing CpG methylation changes genome wide throughout the formation of the three germ-layer, in particular loss of CpG methylation, one can identify the so called DMRs (Fig. 8). These regions reproducibly lose CpG methylation over their initial pluripotent state when differentiating into a certain germ-layer[21, 29]. Looking at those DMRs one can note regions that are not very germ-layer specific (Fig. 8 small dots). Nevertheless, many of them are highly specific for their germ-layer with at least an average CpG methylation decrease of 40% and more. Interestingly, these specific DMRs are mostly small in length and low in CpG count with a few exceptions. Most interestingly, one DMR seems outstanding in length and specificity, a DMR at the *SOX17* locus. This DMR only appears in definitive endoderm and is not seen in ESCs or in any of the other germ-layer (Fig. 9 dashed line).

Since *SOX17* expression is specifically activated in definitive endoderm and not present in any of the other germ-layer or pluripotent ESCs[21, 29], many interesting questions arise in terms of function of the DMR, its impact on *SOX17* gene regulation, the extended epigenetic landscape within the DMR, the presents of regulatory elements and the 3D chromosome architecture of the entire locus throughout the formation of definitive endoderm, which so far has not deeply investigated in human early development. The study of Tsankov A. et al. gives



**Fig. 8 DMRs across the three germ-layers.** Depicted are regions in the human genome that lose methylation upon ESC (PSC) differentiation throughout the formation of ectoderm (EC), mesoderm (ME), or endoderm (EN). DMR specificity for a certain germ-layer is indicated by the dot-size while color depicts the differentiation path. Color intensity indicates the percentage loss of CpG methylation. Note the size of the *SOX17*-DMR and its high specificity to endoderm. (Newly generated figure from Tsankov A. et al., 2015)

initial hints in terms of transcriptional activity at the DMR by gain of H3K4me3 histone marks and the occupancy of several endoderm associated TFs as e.g. EOMES, FOXA1, FOXA2, GATA4, GATA6, SOX17 and OTX2, accompanied by a loss of NANOG binding specifically towards definitive endoderm. But still, several questions remain as e.g. functional questions for the importance of locus-regulation by *SOX17*-DMR and its presence, tissue specificity for definitive endoderm, structural questions of *cis*-regulatory element (CRE) identity and correlation of *SOX17*-DMR activity, and finally the expression of *SOX17* that need to be further investigated in order to understand *SOX17* gene regulation at the locus level during definitive endoderm formation.



**Fig. 9 DNA methylation landscape of the *SOX17* locus.** Depicted are individual CpGs (dots) in each of the three germ-layer (Ectoderm (EC), mesoderm (ME) and endoderm (EN)) and undifferentiated stem cells (ESC). CpG methylation frequencies are given from 0-100% methylation. Note the endoderm specific DMR approx. 230 kb upstream of *SOX17*. (Adapted and modified figure from Tsankov A. et al., 2015)

### 1.5.2 *SOX17* and its diverse functions in development and disease

Within the human gastrula *SOX* genes are expressed all over the embryo in certain pattern. [17] *SOX7* and *SOX17* – Members of the *SOX-F* group – are expressed in cells of the epiblast and hemogenic progenitors. Interestingly, *SOX17* is additionally highly abundant and exclusively expressed in the endoderm cluster of the human gastrula [17]. *SOX18* however is entirely absent throughout the embryo [17]. In adult tissues *SOX-F* members are present in various tissue types, as e.g. *SOX7* with highest expression levels in vaginal and esophageal tissue. *SOX17* expression appears to be most highly in adipose and breast tissue while *SOX18* is most highly expressed in heart muscle and adipose tissue [192].

Despite the broad expression profile across several somatic tissue types [192], in early human embryonic development during gastrulation, *SOX17* expression is most highly abundant in definitive endoderm [17]. *In vitro* studies of human ESCs have revealed distinct roles for *SOX7* and *SOX17*, whereas overexpression of *SOX7* led to development of extraembryonic progeny while *SOX17* overexpression generates definitive endoderm [193]. Development of *SOX17* knock-out mice (*SOX17*-null mutants) has revealed tremendous tissue expansion defects in gut endoderm of the early mid- and hindgut and late foregut development, after neural plate

stage[32]. *SOX17*-null mice develop only until day 10.5 days post coitum (dpc) and live homozygous offspring cannot be recovered after that[32]. In *Xenopus laevis* gastrulae, molecular studies revealed many direct transcriptional targets of Sox17, including *Foxa1* and *Foxa2*. It has been shown that  $\beta$ -Catenin, physically interacts with Sox17 and potentiates its transcriptional activation of target genes. Also, depletion of  $\beta$ -catenin from embryos resulted in a repression of Sox17 target genes. Interestingly, it was also found a motif in the C terminus of Sox17 itself, which is conserved in all the SoxF subfamily Sox proteins, and required for the ability of Sox17 to both transactivate target genes and bind  $\beta$ -catenin.[35] These results were confirmed by a recent study where Sox17 has been shown to functionally interact with the canonical Wnt pathway of *Xenopus laevis* gastrulae to specify and pattern the endoderm while repressing alternative mesoderm fates. In this study, Sox17 and  $\beta$ -Catenin were observed to co-occupy hundreds of key enhancers. It was found instances in which Sox17 and  $\beta$ -catenin synergistically activated transcription independent of Tcfs, whereas on other enhancers, Sox17 repressed  $\beta$ -Catenin/Tcf-mediated transcription to spatially restrict gene expression domains.[33]

Many more studies reported the relation of SOX17 with the Wnt signaling pathway. In cancer, SOX17 is mainly described as a tumor-suppressor[30, 194, 195] since it negatively regulates WNT / $\beta$ -Catenin which is one of many crucial signaling pathways in tumorigenesis and progression[196] of solid and liquid tumor types. In cervical cancer for instance SOX17 restrained the proliferation and tumor formation by down-regulating the activity of the Wnt / $\beta$ -Catenin signaling pathway via trans-suppression of  $\beta$ -Catenin.[197] In endometrial cancer (EC) SOX17 is proposed as a marker for beneficial outcome by inhibiting EC cell migration through the inactivation of the Wnt/ $\beta$ -Catenin driven epithelial-to-mesenchymal transition (EMT) axis.[198]

To understand how tumor-suppressor genes as e.g. SOX17 may be utilized to control disease one needs to understand the molecular interactions and mode of actions of these molecules. However, it is also important to gain insights in how these genes are controlled in certain biological contexts. Since only little is known about the epigenetic and nothing about the genetic control of SOX17 during the formation of definitive endoderm, further investigations may be helpful and translatable to the disease context. Interesting avenues may be explored as e.g. the dynamics of the 3D and chromatin architecture and the identification and functional validation of CREs (enhancer and promoter landscape) within the *SOX17* locus. Therefore, perturbation studies by CRISPR/Cas9 based approaches in combination with *in vitro* definitive endoderm differentiation assays may be most suitable to answer these questions and gain insights into the mechanistic regulation of the human *SOX17* locus.

As outlined throughout this entire chapter, gene regulation in development is controlled by multiple key-players on the DNA, RNA and protein level and their complex interplay

coordinates tight control of developmental genes to guarantee spatiotemporal precision in multicellular organisms. Key aspects here are the proximity of tissue-specific CREs and gene promoters facilitated by TFs and coactivators, transcribed ncRNAs, epigenetic regulators and the complex interplay of all these factors in gene-control. Which role these individual factors play in an individual perspective has only begun to be understood and is still subject of many currently ongoing studies, to gain insights into gene control.

## 2 AIMS OF THE THESIS

SOX17 is a key developmental regulator particularly important for the emergence and maintenance of definitive endoderm in the early human embryo. Its spatiotemporally controlled expression in definitive endoderm is hypothesized to be controlled by its 230 kb upstream differentially methylated region (DMR). The *SOX17* locus including its DMR, is embedded in a gene-desert contained in its own CCCTC-binding factor (CTCF) loop domain, which is further a part of a topologically associating domain (TAD). Gene control by 3-dimensional (3D) chromatin architecture, especially the role of CTCF-boundaries at loop-domains in development, is controversial and only partially understood, as exemplified studies of single loci show. The first aim of this thesis was to identify the role of the *SOX17* loop-domain architecture – with the focus on CTCF boundaries – and its emerging influence on *SOX17* gene control. Hence, I tested the impact of the CTCF loop-domain at the *SOX17* locus in a first, published study by Wu H.J. and Landshammer A. et al. Therefore, I utilized a loss of function approach, deleting a distinct CTCF loop-domain boundary and investigated its importance for loop-architecture and its relevance for gene control of *SOX17* in iPSC derived definitive endoderm. My goal was to characterize potentially resulting developmental phenotypes and connect them to the 3D chromatin architectural based regulation of *SOX17* during the formation of definitive endoderm. Many developmental regulators are located within gene-deserts, somehow “isolated” from other genes and their genetic environment. With this model system thus, I also wanted to test the general notion of developmental regulators being isolated in CTCF loop-domains and challenge their importance and influence on gene-control. Besides 3D chromatin architecture, spatiotemporal gene-control in development is ultimately facilitated by *cis*-regulatory elements (CREs), as shown for many developmental regulators. The second aim of this thesis was to dissect the *SOX17*-DMR in precise detail, within a second, yet unpublished study by Landshammer A. and Bolondi A. et al. It was of interest to determine the epigenetic identity of the DMRs’ potential *cis*-regulatory elements and to study their functional relevance and impact on *SOX17* gene-control. Therefore, I identified two CREs within the DMR, described their identity by epigenetic profiling and tested their regulatory potency in iPSC derived definitive endoderm utilizing CRE-LOF. In the course of this study, I also identified a novel long non-coding RNA (lncRNA) locus within the *SOX17* loop-domain. To prove its lncRNA identity, I carried out in-depth characterization of the RNA-molecule, studied its biogenesis and probed its cellular localization. Since lncRNAs and their active transcription often participate in developmental gene-control, I utilized two different LOF approaches to test for either RNA or transcription, potentially involved in *SOX17 cis*-regulation. Further, I explored the lncRNA’s general relevance for the formation of definitive endoderm, ultimately validating developmental potency of lncRNA lacking definitive endoderm in iPSC derived pancreatic progenitors.



### 3 THESIS CONTRIBUTIONS

#### 3.1 Publication: Topological isolation of developmental regulators in mammalian genomes.

Authors

Hua-Jun Wu\*, **Alexandro Landshammer\***, K. Stamenova, Adriano Bolondi, Helene Kretzmer, Alexander Meissner & Franziska Michor

\* Wu and Landshammer contributed equally to the study

Published in *Nature Communications*, August 2021, 12:4897.

doi: [10.1038/s41467-021-24951-7](https://doi.org/10.1038/s41467-021-24951-7)

Personal contribution

I designed and performed experiments, collected, analyzed, and interpreted the data, prepared the visualization of the work, and arranged the figures, and wrote the corresponding text. In detail, I performed experiments that contributed to the following figures of the publication ([s. below under 8. Appendices, 8.4. Attachment](#)):

Fig. 4d-f, Fig. 5, Supplementary Fig. 5a-d, f-i, Supplementary Fig. 6b-d, Supplementary Fig. 7, Supplementary Fig. 8.

The corresponding figures in this thesis I contributed to are Fig. 15-20.

### **3.2 Unpublished: Discovery and characterization of *LNCsox17* as an essential regulator in endoderm formation.**

#### Authors

**Alexandro Landshammer\***, Adriano Bolondi\*, Helene Kretzmer, Christian Much, René Buschow, Alina Rose, Hua-Jun Wu, Sebastian Mackowiak, Bjoern Braendl, Pay Giesselmann, Rosaria Tornisiello, Krishna Mohan Parsi, Jack Huey, Thorsten Mielke, David Meierhofer, René Maehr, Denes Hnisz, Franziska Michor, John L. Rinn & Alexander Meissner.

\* Landshammer and Bolondi contributed equally to the study

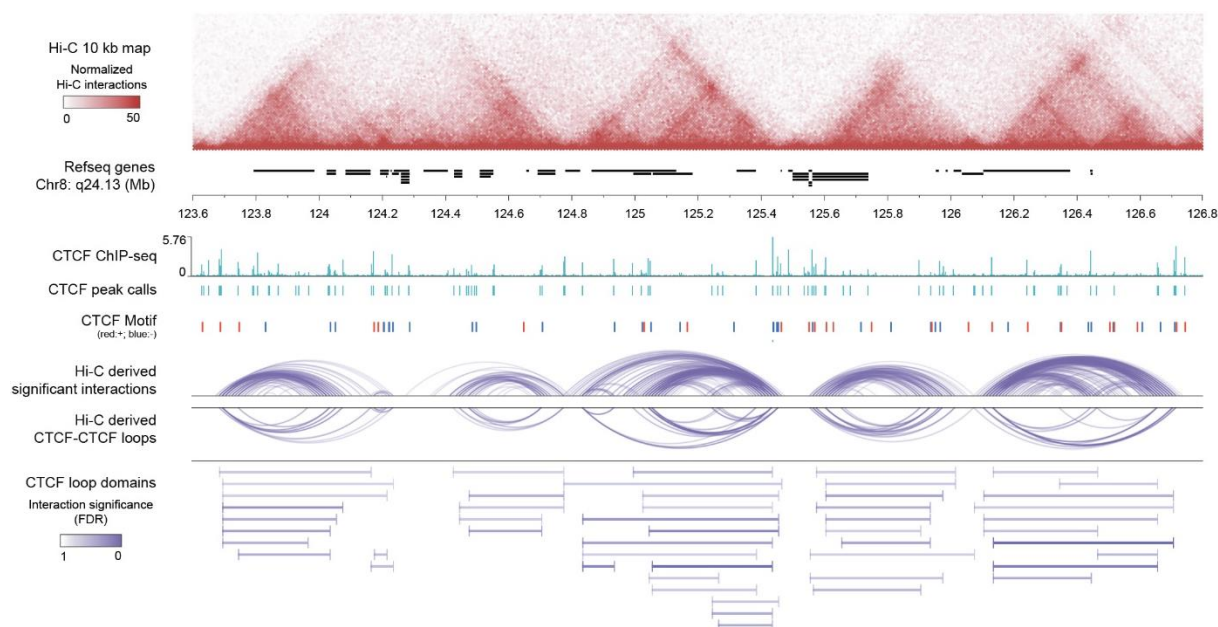
#### Personal contribution

I designed and performed experiments, collected, analyzed, and interpreted the data, prepared the visualization of the work, and arranged the figures, and wrote the corresponding text. In detail, I performed experiments that contributed to figures in this thesis namely Fig. 21-33.

## 4 RESULTS

### 4.1 Identification of single gene loop domains and topologically isolated genes (TIGs)

Gene expression and the precise control of mammalian genomes is facilitated by nuclear organization and epigenetic regulatory mechanisms. Insulated chromosomal territories are important for the regulatory control of genes but the biological relevance of their boundaries in developmental processes are highly complex and yet, remain partially understood. In our previous study by Wu H.J. and Landshammer et al. [91], we generated deeply sequenced Hi-C data from human pluripotent stem cells (hPSCs) and their three germ-layer derivatives. We further processed these computationally and identified three general sub-types of CTCF loop domains with highly conserved boundary CTCF sites. We found that one of those domain types contain only a single protein-coding gene (PCG) and are spanned by highly conserved CTCF sites, significantly enriched for developmental regulators. To our surprise we found that the endodermal transcription factor (TF) *SOX17*, is among this group of so called topologically isolated genes (TIGs). Therefore, we used *SOX17* as a TIG-exemplary locus in our study to show that perturbation of such a boundary leads to deregulated distal enhancer-promoter interactions, interfering with the definitive endoderm *in vitro* differentiation model system.

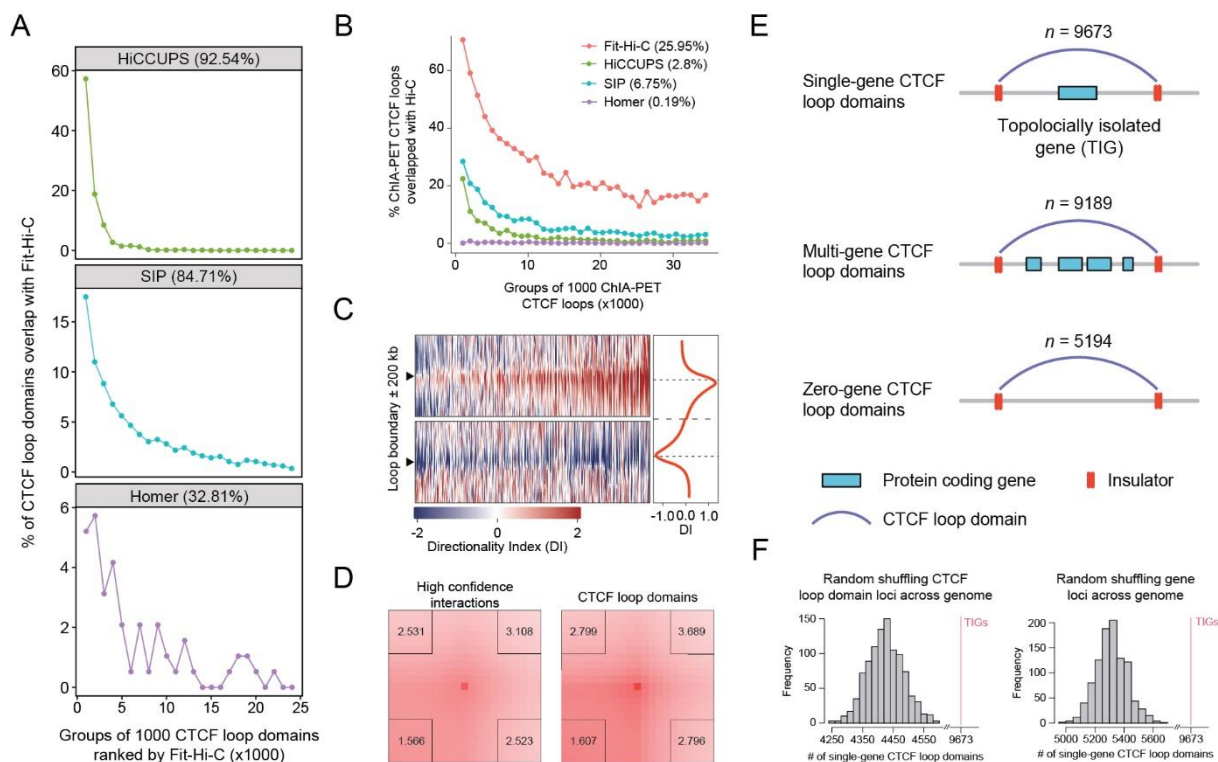


**Fig. 10 Identification of topologically insulated CTCF loop domains in HUES64 ESC genomes.** Depicted are from top to down, a normalized Hi-C interaction heatmap of HUES64 (25 kb resolution per pixel), Refseq genes at chromosome region 8q24.13, a normalized CTCF-ChIP sequencing profile of HUES64 and its resulting called CTCF peaks and respective CTCF consensus motifs at the locus denoted by red (forward orientation) and blue (reverse orientation). Further, resulting high confidence Hi-C interactions and CTCF-loops displayed as arcs and the final resulting CTCF-loop domains. Interaction significance is given as false-discovery-rate (FDR) calculated from Hi-C data. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J. and Stamenova E. contributed to this figure)

To investigate 3D chromatin architecture and to study its role in human pluripotent stem cells (hPSC), we performed Hi-C sequencing[99] of the human embryonic stem cell line (ES) HUES64. 1.05 billion uniquely mapped paired-end reads were generated. Applying Fit-Hi-C[199] on these data, we could identify 231,970 high confidence interactions in the genomic range of 20 kb – 2 Mb. Mapping the regarding Hi-C interactions to the union of CTCF motifs and CTCF ChIP sequencing peaks in HUES64 hPSCs, basically the collection of CTCF sites, we found 37,428 significant CTCF-CTCF loops. To limit redundancy, loops in proximity to each other were further clustered and merged, resulting in a total of 24,056 CTCF loop domains with a median length of 304 kb (Fig. 10, Fig. 11D).

To assure that the observed domain calling results are valid and called technically correct, we compared different Hi-C loop detection methods[200-202], all independently identifying Hi-C loops. We observe consistent and high levels of agreement between Fit-Hi-C and HiCCUPS (92.45%) or SIP (84.71%) (Fig. 11A). To test the overlap between Fit-Hi-C actual CTCF-CTCF loop domains with proximity ligated genomic interactions following CTCF antibody-enrichment, we found Fit-Hi-C calls to overlap highly with insulated neighborhoods identified by Cohesin ChIA-PET data in primed human ES cells[203], compared to other calling methods (Fig. 11B). To demonstrate a topological insulation function for all CTCF loop domains including their surrounding regions, we further calculated a so-called directionality index[100] (DI) (Fig. 11C). The DI provides a quantification of degree for an upstream and downstream bias of distinct genomic regions, which helped us to identify that many CTCF loop domains contain only one protein-coding gene (PCG) (~40%,  $n = 9,673$ ) compared to domains containing multiple genes (~38%,  $n = 9,189$ ) or no genes (~22%,  $n = 5,194$ ) (Fig. 11E). To challenge the observed enrichment of single protein-coding genes within many CTCF loop-domains we compared the representations of genes over randomly shuffling either domains or genes across the genome and finds a significant overrepresentation (permutation test  $p < 0.001$  Fig. 11F). We decided to term these CTCF loop domains as single-gene loop domains, and the genes they contain as topologically isolated genes (TIGs) (Fig. 11E).

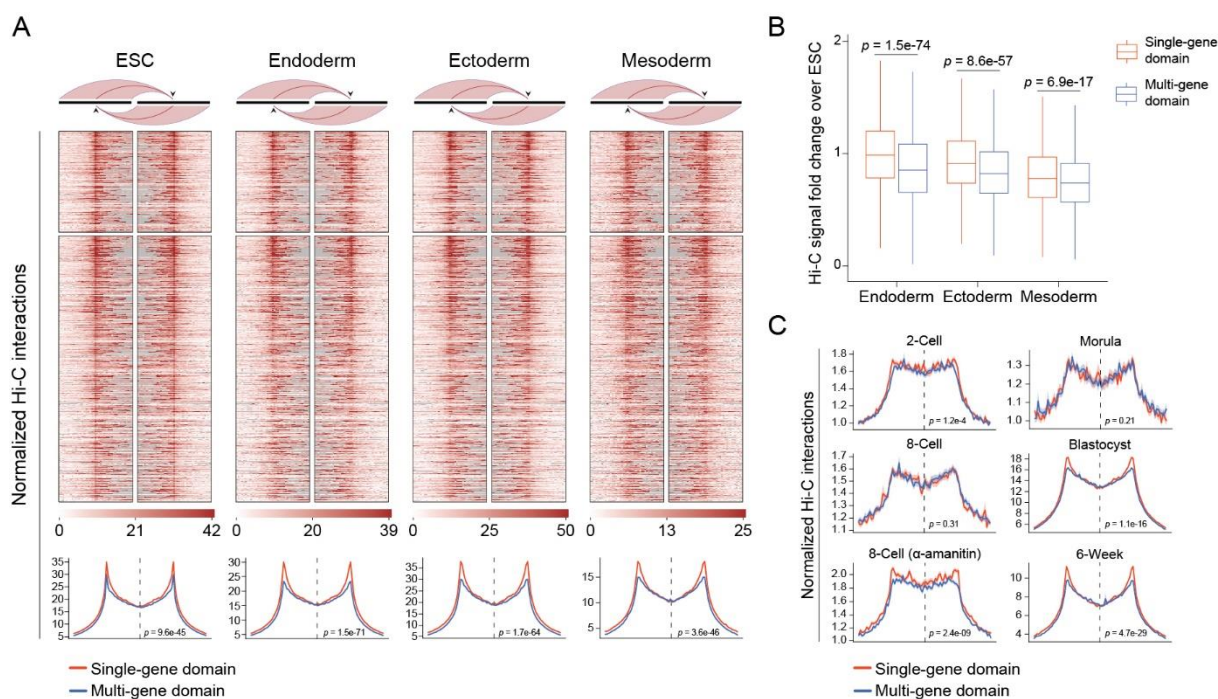
The identification of TIG domains raised a variety of interesting questions regarding their boundaries such as insulation and gene-regulation function, phylogeny, conservation, and stability in early development and across different tissues, hence we carried out further analysis trying to answer these and to characterize TIGs and their CTCF boundaries further.



**Fig. 11 CTF single gene loop domains and their topologically isolated genes (TIGs).** (A) Comparisons of CTF loop domains identified by Fit-Hi-C and other methods. The percent of CTF loop domains overlapping between Fit-Hi-C and HiCCUPS, SIP, Homer in HUES64 Hi-C data are plotted for loops ranked from the most significant to the least significant by Fit-Hi-C. Each point represents 1000 loops. The total percentage of CTF loop domains overlapping for each method is shown in parentheses on top of each subplot. The CTF loop domains are defined to be exactly the same without any shift in this analysis. (B) Percentage of insulated neighborhoods identified by ChIA-PET (x-axis) overlapping with CTF loop domains obtained by different methods in Hi-C data (y-axis) are depicted for insulated neighborhoods ranked by ChIA-PET data from the most significant to the least significant. Total percentages of insulated neighborhoods overlapping with different Hi-C methods are indicated in the legend. Each point represents 1000 loops. Insulated neighborhoods in ChIA-PET were obtained from the supplementary table of the original paper (Ji X. et al., 2016). (C) Directionality Indices (DI) obtained from 10 kb Hi-C map in HUES64 cells by Homer are plotted for the surrounding regions (200 kb upstream and downstream) of the left and right CTF loop domain boundaries to depict their insulation function. The left heatmap shows the surrounding regions of the left boundary, and the right heatmap shows the surrounding regions of the right boundary. Each row represents one CTF loop domain. Black arrow heads represent the left and right boundaries. The average profiles are shown on top of the heatmap. Higher signal in the left heatmap represents entering an insulated region, and lower signal in the right heatmap represents exiting an insulated region. (D) Genome wide data of 231,970 high confidence interactions (left) and 24,056 CTF loop domains (right) called from HUES64 Hi-C data and depicted as aggregate peak analysis (APA) plot. APA scores (numbers in the corners) is the ratio of the number of contacts in the central bin to the mean number of contacts in the corner. (E) Scheme of single-gene CTF loop domains, multi-gene domains, and zero-gene domains. The numbers of domains in each group within HUES64 are displayed on top of each plot. (F) Distribution of the number for single-gene domains in the human genome by random shuffling gene loci across the genome (top). Distribution of the number for single-gene domains in the human genome by randomly shuffling the domains across the genome (bottom). The red line indicates the observed number of single-gene domains across the genome. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J. and Stamenova E. contributed to this figure)

## 4.2 Characterization of CTCF loop domains and their boundary elements

To explore stability of CTCF loop domain boundaries in the three germ-layer and compare them to hPSCs, we utilized *in vitro* of the HUES64 line to generate Hi-C data from derived ectoderm (EC), mesoderm (ME) and definitive endoderm (EN). Initial analysis revealed CTCF loop domain boundaries to be preserved during ES cell differentiation, however a more suitable approach to analyze large numbers of boundaries at the same time was necessary. Hence,



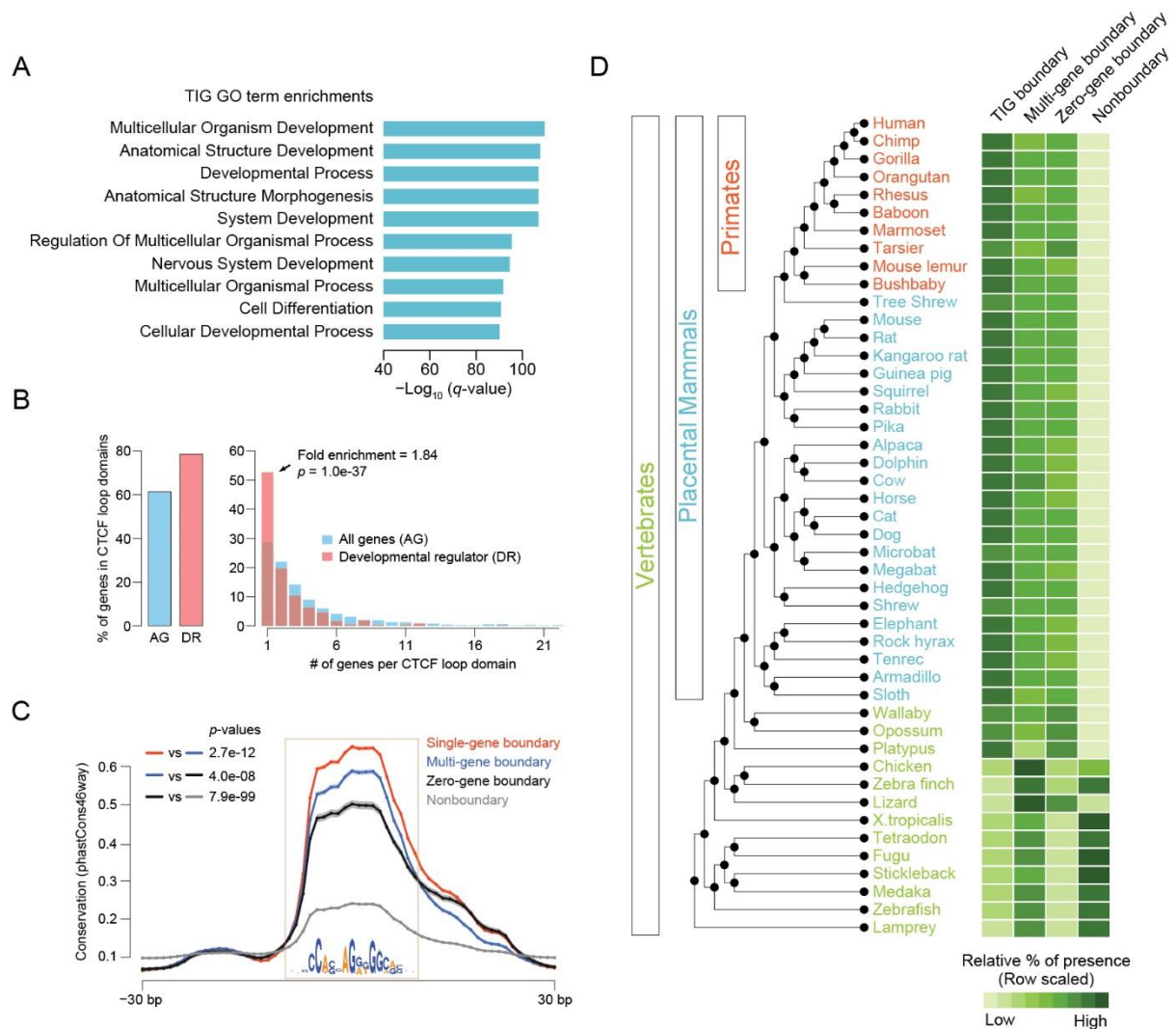
**Fig. 12 CTCF loop domain establishment and stability throughout early development.** (A) Top panel indicates boundary-anchored virtual 4C heatmaps of domain boundaries identified from HUES64 Hi-C data underlying contact maps. Depicted are the normalized Hi-C interactions. Bottom panel shows boundary-anchored virtual 4C average profiles of domain boundaries in HUES64 derivatives. Domain boundary locations were identified in HUES64 Hi-C data. Depicted are normalized Hi-C interactions. The dotted line separates left and right boundary regions, which represent regions in the left and right heatmap of the top panel. Average signals across all boundaries are shown by the shaded area indicating the standard error. Two-sided Wilcoxon test was used to determine significance levels of boundary-to-boundary interactions between the two groups. Data are presented as mean values  $\pm$  SE. (B) Hi-C signal fold-change of boundary-to-boundary interactions in germ-layer derivatives over HUES64 cells. Shown is the two-sided Wilcoxon test  $p$ -value. Hi-C signals were normalized by library size in individual samples prior to the analysis. The box indicates the interquartile range (IQR), the line inside the box shows the median, and whiskers show the locations of either  $1.5 \times$  IQR above the third quartile or  $1.5 \times$  IQR below the first quartile,  $n = 3,310$  boundary-to-boundary interactions for single-gene domains,  $n = 8,729$  boundary-to-boundary interactions for multi-gene domains. (C) Boundary-anchored virtual 4C average profiles of the domain boundaries at the 2-cell, 8-cell, 8-cell treated with  $\alpha$ -amanitin, morula, blastocyst stages, and 6-week embryos. The locations of domain boundaries were identified in HUES64 Hi-C data. CTCF expression is inhibited under  $\alpha$ -amanitin treatment at the 8-cell stage. Shown are the normalized Hi-C interactions. The dotted line separates the left and right boundary regions. Average signals across all boundaries are depicted by the shaded area indicating the standard error. Two-sided Wilcoxon test was used to determine the significance level of boundary-to-boundary interactions between the two groups. Data are presented as mean values  $\pm$  SE. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J. and Stamenova E. contributed to this figure)

boundary-anchored virtual circular chromatin confirmation capture (4C) plots were used to visualize contact interactions between all boundary pairs within a sample. In this approach two heatmaps are used in which the left heatmap represents the Hi-C interactions from the surrounding genomic regions of left-to-right boundaries (setting left boundaries as 4C viewpoints) and the right heatmap Hi-C interactions from right-to-left boundary anchored surrounding regions (setting right boundaries as 4C viewpoints) (Fig. 12A upper panel). As soon as there are physical interactions observed between the two boundaries within a sample, the plot exhibits a high intensity in the center of both the heatmaps but not the surrounding regions. Doing so, CTCF loop domain boundaries were observed to be largely preserved throughout *in vitro* three germ-layer differentiation as depicted by normalized virtual 4C contacts (Fig. 12A upper panel).

To profile different domain types within a sample, heatmaps were aggregated by column to generate a profile plot representing the average and standard deviation of the signal across all boundaries. By the help with this analysis, we found clear peaks in boundary centers of TIGs and multi-gene domains (Fig. 12A lower panel), which highlights that TIG boundaries are preserved more than multi-gene CTCF loop domain boundaries (Fig. 12B).

To sought whether and when during embryonic development hESC CTCF loop domain boundaries are formed, we further analyzed public Hi-C data of 2-cell, 8-cell, morula, blastocyst and 6-week staged human embryos[204]. We found CTCF loop domain boundaries to be established gradually from 8-cell stage on, shortly happening after zygotic genome activation (ZGA)[205] and the induction of CTCF expression[204]. As indicated by the data, stability of these boundaries is maintained from blastocyst stage onwards (Fig. 12C). Blastocyst and 6-week stage TIG boundaries were more pronounced than multi-gene domain boundaries indicated by Hi-C, as observed earlier in hESCs and their three germ-layer derivatives (Fig. 12A,C). Interestingly, this was not the case in earlier stages as e.g. 8-cell and morula, indicating that boundary divergence arises between morula and blastocyst stages which coincides with the initiation of lineage specification[206]. Hi-C data of ZGA inhibited 8-cell stage embryos by  $\alpha$ -amanitin treatment (RNA PolIII inhibitor and CTCF repressor)[207-210] had less influence on TIG boundaries than multi-gene boundaries (Fig. 12C). From these observations we conclude that once formed, TIG boundaries are maintained ZGA and CTCF-expression independent, hence being more stable and robust across developmental processes.

The establishment of TIG boundaries in the early human embryo and their preservation during the three germ-layer formation implies regulatory function and importance during early development. Interestingly, analyzing TIGs further we found an enrichment for diverse developmental processes within the Gene Ontology (GO) database (Fig. 13A). We further defined developmental regulators (DRs) as TFs under the GO term “developmental process”.When performing enrichment analysis of different CTCF loop domain types, based



**Fig. 13 Gene ontology (GO), conservation, and phylogeny of CTCF loop domains.** (A) TIG gene ontology (GO) enriched terms for HUES64. Overall enriched are diverse developmental processes. (B) Gene enrichment analysis across different loop domains. Depicted in the left panel are percentages for all genes (AG) and developmental regulators (DR) located in all loop domains across the genome. Shown in the right panel, percentages for both AG and DR within a certain group of gene-number per loop domain. Single-gene loop domains are highly enriched for DRs of AGs.  $p$ -value is calculated by two-sided Fisher's exact test. (C) Evolutionary conservation analysis of CTCF consensus motifs at boundaries of different domain types. Nonboundary CTCF motifs represent the motifs that are outside of any domain boundaries. Given within the box, the motif region, and the motif sequence. Displayed is the average conservation score (phastCons46way) across placental mammals and all boundary regions. The shaded area indicates the standard error.  $p$ -value were tested by the two-sided Wilcoxon test for the given motif region. Data are presented as mean values  $\pm$  SE. (D) Analysis of the presence or absence of human CTCF motifs among the 45 vertebrates. CTCF motifs were grouped based on the CTCF loop domain classification in HUES64. The relative proportion of present motifs in each group per species were calculated and shown in the right panel. Specifically, the absolute proportion of present motifs was calculated and row Z-scored to obtain the relative proportion of present motifs. The phylogenetic tree is obtained from UCSC genome browser. Relative TIG boundary presence arises with marsupials. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J. and Stamenova E. contributed to this figure)

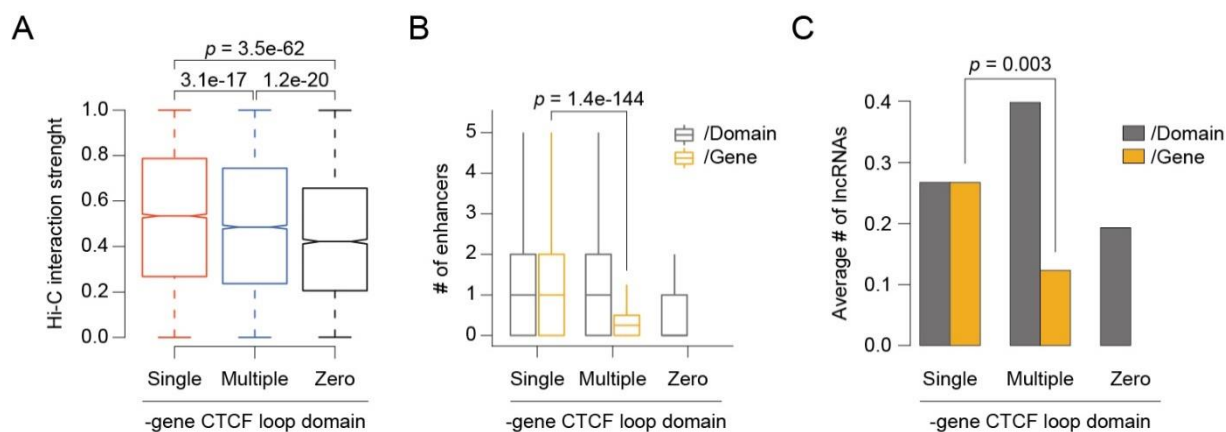
on domain insulated number of genes, we found especially these genes to be enriched within single gene CTCF loop domains (Fig. 13B). Interestingly, this enrichment is not found for



domains insulating multiple genes, which led to the motivation to sought whether the insulation function for these boundaries is functionally important.

To explore this further, we analyzed the conservation-extend of these boundaries across different species and identified CTCF boundary motifs of single-gene loop domains to be highly conserved with the emergence of marsupials but especially within placental mammals (Fig. 13C,D). This may suggest that CTCF motifs of single-gene loop domains are functionally important elements undergoing natural selection. CTCF motifs of boundaries insulating zero or multi-gene domains instead showed a sequentially decreasing conservation score instead, while outside boundary CTCF motifs were generally not conserved, or to a much lower extend than all others (Fig. 13C,D). Since single-gene CTCF loop domains enrich for DRs, together these analyses suggested CTCF boundaries of those domains to be functionally important.

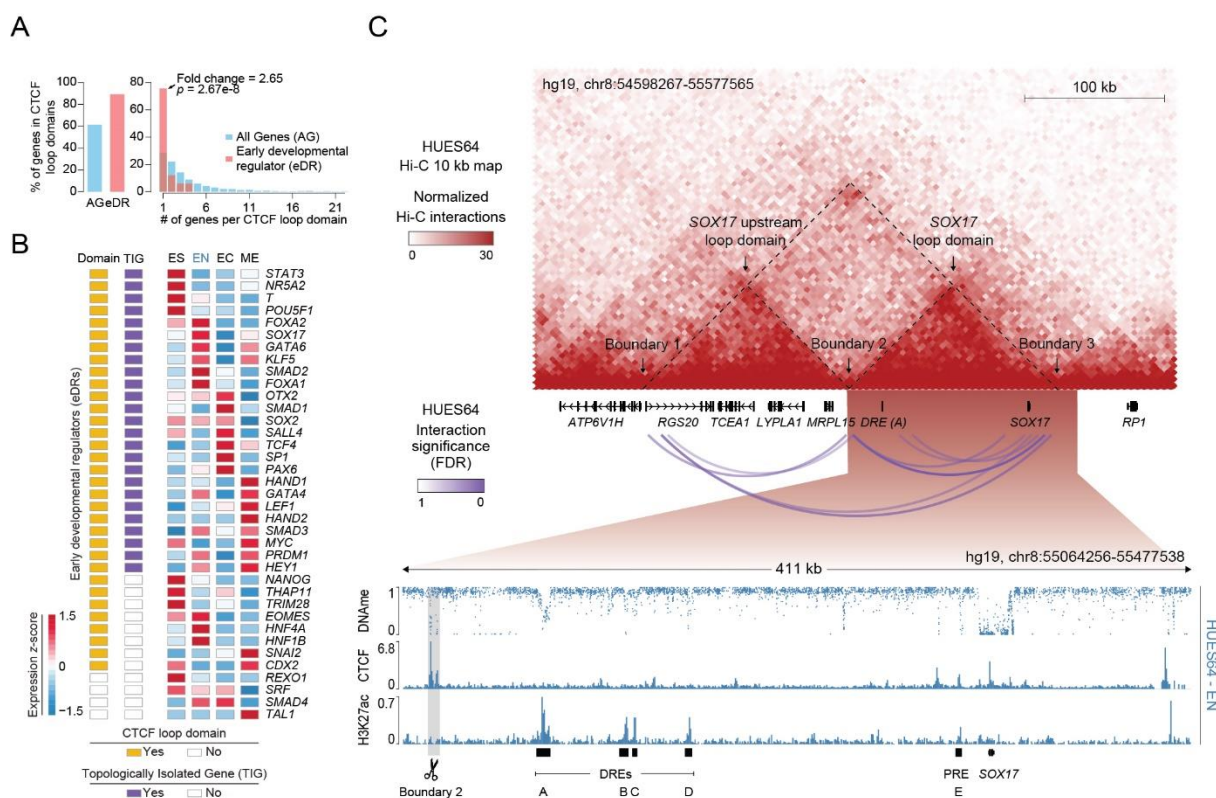
Based on Hi-C data, we also overserved boundaries of single-gene loop domains to interact stronger than others with a similar average signal intensity for TIG and multi-gene domain boundaries (Fig. 14A). Interestingly these domains also contain more *cis*-regulatory elements (CREs) per gene as e.g., enhancers and long non-coding RNAs (lncRNAs) than the other types of domains (Fig. 14B,C). Altogether, these results imply TIGs to be functionally important underlined by the fact of DR enrichment, which was further supported by the conservation patterns across mammals.



**Fig. 14 CTCF loop domain strength and CRE abundance. (A)** Hi-C interaction strength of different types of domain types. The significance ( $-\log_{10} p$ -value) of Hi-C interaction is rank normalized and displayed on the y-axis with “1” represents most and “0” the least significant interactions. **(B)** Number (#) of enhancers per domain (or per gene) within different types of domains. In grey represented the number of features per domain and in orange the number of features per gene. **(A,B)** The box displays the interquartile range (IQR), the line inside the box the median, and whiskers show the locations of either  $1.5 \times$  IQR above the third quartile or  $1.5 \times$  IQR below the first quartile,  $n = 9673$  single-gene domains,  $n = 9189$  multiple-gene domains,  $n = 5194$  zero-gene domains. **(C)** The average number of long non-coding RNAs (lncRNAs) within different types of domains. **(A-C)**  $p$ -values were calculated by two-sided Wilcoxon test. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J. and Stamenova E. contributed to this figure)

### 4.3 Genetic dissection and characterization of the *SOX17* loop domain

To test whether genetic ablation of TIG boundaries would have an effect on gene-regulation of insulated genes throughout hESC differentiation, we refined the list of DRs curated from the HUES64 cell line to specifically include early developmental regulators (eDRs). eDRs display a stronger enrichment in both CTCF loop domains and TIGs than other DRs (Fig. 13B and Fig. 15A). We found most eDRs located within CTCF loop domains (89%, 33/37), 67% to be located in single-gene loop domains (TIGs) (25/37), 22% in multi-gene loop domains (8/37) and 11% in CTCF loop domain free regions (4/37) (Fig. 15B). To enable a functional characterization of a representative TIG whose gene is specifically activated during gastrulation, we decided for



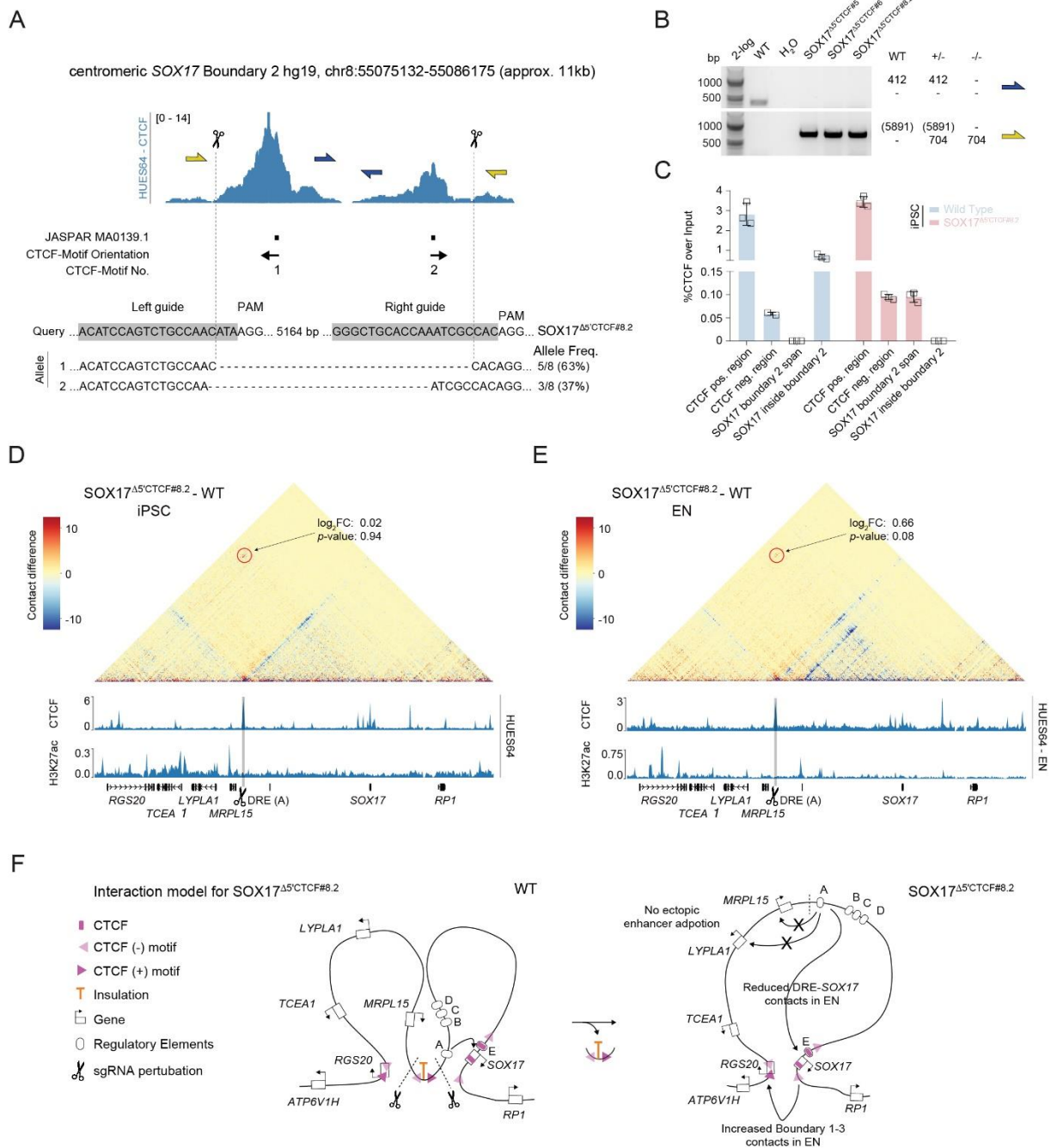
**Fig. 15 Overview of the *SOX17* CTCF loop domain at a locus resolution. (A)** Enrichment of early developmental regulators in single-gene domains. The left panel represents the percent of all genes (AG) or early developmental regulators (eDR) located within domains. The right panel represents the percent of AG or eDR located within domains with increasing number of protein coding genes.  $p$ -values calculated by two-sided Fisher's exact test. **(B)** Heatmap of early developmental regulators (eDRs) displaying information on CTCF loop domains, TIGs, and expression in embryonic stem cell (ES) three germ-layer differentiation. The RPKM (Reads per kilobase per million mapped reads) value of gene expression in embryonic stem cells (ES), definitive endoderm (EN), ectoderm (EC), and mesoderm (ME) were row z-scored. **(C)** Top panel shows a normalized Hi-C interaction map of the *SOX17* locus as a representative TIG at hg19, chr8:54598267-55577565. HUES64 CTCF loop domains are displayed as arcs below. Bottom panel highlights HUES64 derived EN WGBS sequencing, CTCF, and H3K27ac ChIP-seq profiles. Putative *SOX17* distal regulatory elements (DRE) and proximal regulatory elements (PRE) are depicted by black bars and given capital letters. The deleted centromeric *SOX17* boundary (Boundary 2) is highlighted in grey and marked by a scissor. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Wu H.J., Landshammer A. and Stamenova E. contributed to this figure)

the *SOX17* locus as being isolated by strong boundaries (Boundary interaction strength adjusted  $p$ -value =  $3.5e-9$  based on Hi-C data). *SOX17* encodes for a member of the SOX (SRY-related HMG-box) superfamily of TFs that is specifically induced in definitive endoderm differentiation by potential distal regulatory elements (DREs)[29] (Fig. 15B,C).

To interrogate TIG boundary perturbation effects at the *SOX17* locus, we designed sgRNAs flanking the 5' centromeric boundary of the *SOX17* CTCF loop domain at Boundary 2, roughly 300 kb upstream of the locus (Fig. 15C and Fig. 16A). We generated three independent homozygous *SOX17*<sup>Δ5'CTCF</sup> knock-out (KO) clones (Fig. 16A,B) in the female induced pluripotent stem cell (iPSC) line ZIP13K2[211]. Besides human iPSCs and ESCs being transcriptomically[28] highly similar and sharing an almost identical 3D chromatin architecture[212], using iPSCs has various benefits such as sharing and exchanging material and data across labs. It has also been shown that the CTCF occupancy and Hi-C boundary strength at the *SOX17* locus are highly similar between human iPSCs and ESCs[91], hence iPSCs are a valid model to study boundary perturbations at the *SOX17* locus.

Utilizing one of the three independent iPSC KO clones (*SOX17*<sup>Δ5'CTCF#8.2</sup>) further, we confirmed a 5 kb deletion including two CTCF peaks (Fig. 15C, Fig. 16A,B) and the regarding loss of CTCF occupancy at the corresponding *SOX17* Boundary 2 compared to control regions by CTCF ChIP qRT-PCR (Fig. 16C). Capture Hi-C of the locus reveals loss of the corresponding Boundary 2 interactions in *SOX17*<sup>Δ5'CTCF#8.2</sup> human iPSCs without any significant interaction change at upstream/downstream Boundary 1-3 (Fig. 16D red circle). When differentiating cells into definitive endoderm, loss of corresponding Boundary 2 interactions is preserved, while Boundary 1-3 interactions increase specifically in *SOX17*<sup>Δ5'CTCF#8.2</sup> KO cells (Fig. 16E red circle). Moreover, we observed a significant loss of intra-loop interactions (Fig. 16E *SOX17* loop domain) in *SOX17*<sup>Δ5'CTCF#8.2</sup> KO iPSCs as well as a loss of endoderm-specific enhancer contacts between the *SOX17* promoter and its distal regulatory elements (DRE A-D) (Fig. 16E). Without any further evidence of ectopic enhancer hijacking/adoption or alternative enhancer interactions due to perturbation of Boundary 2, we conclude that there is a decreased interaction-frequency of enhancer-gene contacts during the formation of definitive endoderm between the *SOX17* promoter and its tissue-specific enhancer DRE (A) including other DREs (B-D) in *SOX17*<sup>Δ5'CTCF#8.2</sup> KO cells (Fig. 16F).

*SOX17* is a key marker and known as a transcriptional driver and maintenance TF for definitive endoderm[29, 32]. Within development in the early embryo, *SOX17* is often used to identify various endodermal tissues as e.g., primitive, visceral, and definitive endoderm[7]. The combination of transmembrane C-X-C chemokine receptor 4 (CXCR4) and *SOX17* serves as a specific tool to identify definitive endodermal tissue[7, 29]. When utilizing *in vitro* directed diff-

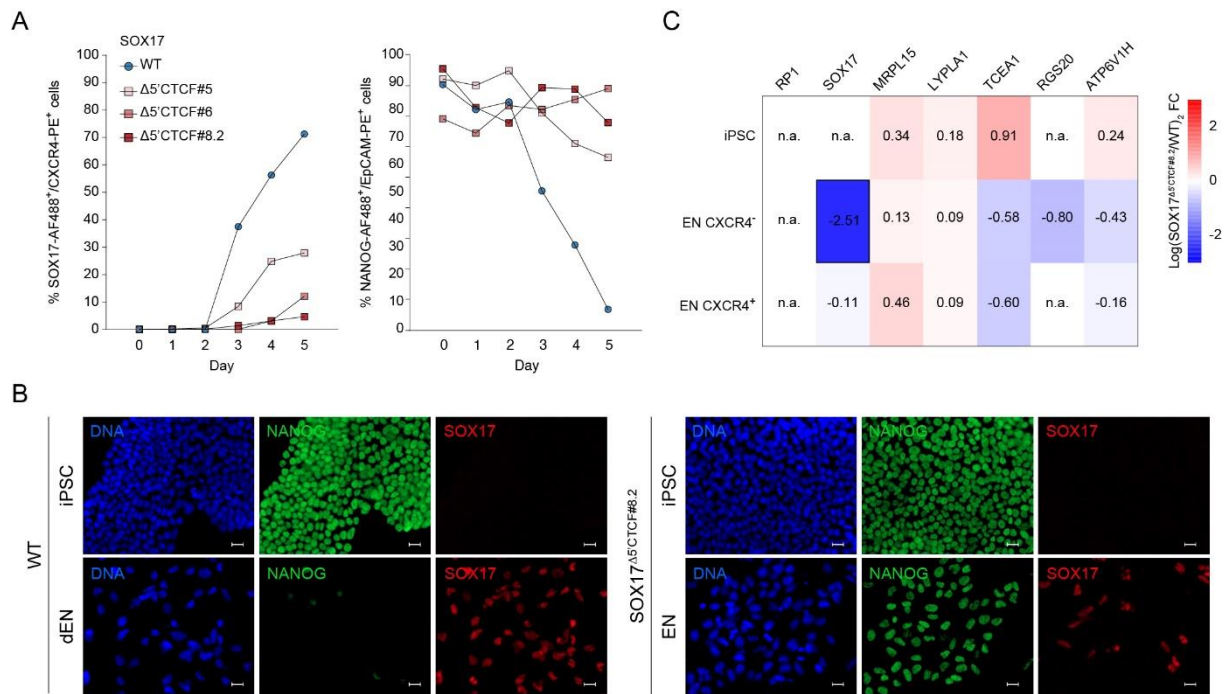


**Fig. 16 Boundary 2 CRISPR/Cas9 perturbation of the *SOX17* CTCF loop domain in detail. (A)** Centromeric *SOX17* Boundary 2 (*SOX17*<sup>Δ5'CTCF</sup>) targeting strategy scheme. Scissors and dashed lines indicate CRISPR/Cas9 target-sites. Representative chromatogram-data including allele-frequency of sanger-sequenced, pJET1.2 cloned PCR-products of clone *SOX17*<sup>Δ5'CTCF#8.2</sup> are depicted below. **(B)** Expected PCR-band pattern for different primer-pairs of centromeric *SOX17* Boundary 2 targeting are indicated by primer color. 1% agarose-gel pictures indicate genotyping PCR band size-separations of respective knockout clone and control gDNA. **(C)** CTCF ChIP-qRT-PCR of *SOX17* Boundary 2 and control regions in undifferentiated *SOX17*<sup>Δ5'CTCF#8.2</sup> or wild-type iPSC. CTCF-enrichment is expressed as percentage of CTCF over input, at the respective site of interest across independent experiments ( $n = 3$ ). Data are presented as mean values  $\pm$  SD. **(D-E)** ChHi-C subtraction maps in iPSCs (**D**) and EN cells (**E**) at the *SOX17* locus. The relative contact difference between the two samples (*SOX17*<sup>Δ5'CTCF#8.2</sup>/wild-type) in either iPSCs or EN cells are shown on top of HUES64 or HUES64 derived EN CTCF and H3K27ac ChIP-seq profiles. Boundary 1+3 contact quantifications are highlighted in red circles. The deleted Boundary 2 is highlighted in grey marked by a scissor. *SOX17*DRE (A) and gene bodies are highlighted in black bars. **(F)** Simplified 2D-model of the *SOX17* Boundary 2 perturbation in wild-type or *SOX17*<sup>Δ5'CTCF</sup> cells. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Landshammer A. and Wu H.J. contributed to this figure)

differentiation conditions to generate definitive endoderm, day 5 terminal differentiations of temporally resolved endoderm revealed a strong reduction in SOX17<sup>+</sup>/CXCR4<sup>+</sup> populations of all SOX17<sup>Δ5'CTCF</sup> isogenic cell lines (on average 4.68-27.95%) compared to wild type iPSCs (71.3%) (Fig. 17A left panel) as observed by FACS analysis. This result was further confirmed by IF stainings (Fig. 17B). To determine if SOX17<sup>Δ5'CTCF</sup> cells would properly exit pluripotency due to differentiation and lose their epithelial character due to epithelial-to-mesenchymal transition (EMT)<sup>194,195</sup>, we tested for pluripotency factor NANOG in combination with the transmembrane glycoprotein Epithelial Cell Adhesion Molecule (EpCAM). As confirmed by IF stainings of day 5 endoderm for SOX17<sup>Δ5'CTCF#8.2</sup> cells (Fig. 17B), temporal FACS analysis revealed reduced NANOG<sup>+</sup>/EpCAM<sup>+</sup> cell populations in wild type iPSCs (16.9%), while cell population numbers remained comparably high over time in all SOX17<sup>Δ5'CTCF</sup> clones (on average 66.7-89.25%) (Fig. 17A right panel). According to this data, we suggest a boundary-dependent deregulation of SOX17 gene-control during the formation of definitive endoderm along with the associated failure to properly exit pluripotency.

To confirm our Hi-C results and test whether the loss of insulation due to Boundary 2 perturbation in clone SOX17<sup>Δ5'CTCF#8.2</sup> would influence upstream gene expression in an endoderm-specific fashion, we examined the absence of a potential enhancer hijacking/adoption scenario for SOX17 DREs (Fig. 15C and Fig. 16F) by in depth transcriptomic characterization. Hence, we FACS-sorted different CXCR4 sub-fractions for RNA sequencing followed by differential gene expression analysis. Doing so, we confirm no deregulation of gene expression for genes within the SOX17 upstream CTCF loop domain (Fig. 17C). Nevertheless, we observe SOX17 expression deregulation only in the minor fraction (on average 4.68 - 27.95%) of CXCR4<sup>+</sup> SOX17<sup>Δ5'CTCF#8.2</sup> cells, which is highly in concordance with previous FACS and IF data (Fig. 17A-C). Taken together, this suggests that SOX17 gene deregulation due to loss of Boundary 2 is not associated with ectopic DRE hijacking/adoption by SOX17 upstream CTCF loop domain related genes.

To deeply investigate transcriptomes of the various differentiated cell populations, we performed principle component analysis (PCA) of the 100 most variable genes across all samples (Fig. 18A,B). SOX17<sup>Δ5'CTCF#8.2</sup> CXCR4<sup>+</sup> and wild type CXCR4<sup>-</sup> cell populations closely clustered together on an endodermal differentiation trajectory roughly between undifferentiated and CXCR4<sup>+</sup> wild type populations (Fig. 18A). Since we were interested in the transcriptomic differences between the respective majority cell populations, we analyzed differentially expressed genes (1,506 genes) of CXCR4<sup>+</sup> wild type and CXCR4<sup>-</sup> SOX17<sup>Δ5'CTCF#8.2</sup> cells using Gene Set Enrichment Analysis (GSEA) for biological processes ( $\log_2FC > 2$ ,  $q$ -value  $< 0.05$ ) and found genes enriched for DNA replication and cell cycle checkpoint in CXCR4<sup>-</sup> SOX17<sup>Δ5'CTCF#8.2</sup> cells (Fig. 18C). Determination of endodermal cell fate propensity is closely connected to the cell cycle<sup>[213]</sup>, which together with the PCA-trajectory suggested that

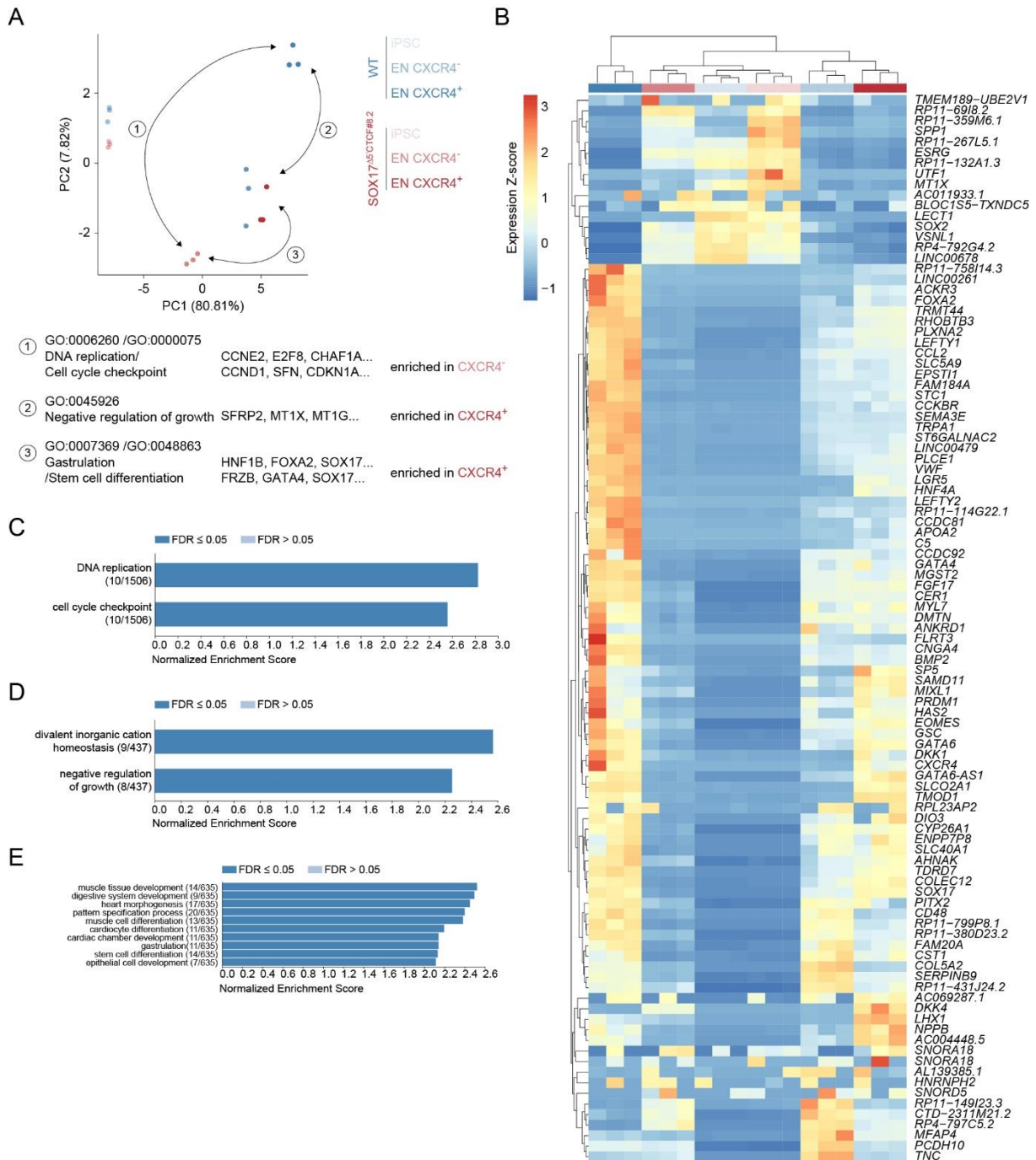


**Fig. 17 SOX17 Boundary 2 perturbation leads SOX17 deregulation without affecting local gene expression.**

**(A)** Fluorescence activated cell sorting (FACS) time-course data of wild-type and SOX17<sup>Δ5CTCF</sup> iPSC during directed differentiation towards definitive endoderm. SOX17 and CXCR4 (CD184) are depicted as percentage SOX17<sup>+</sup>/CXCR4<sup>+</sup> in bulk cell populations. Corresponding NANOG and Ep-CAM (CD326) are depicted as percentage NANOG<sup>+</sup>/Ep-CAM<sup>+</sup> in bulk cell populations. Symbols represent the mean across independent experiments ( $n = 2$ ). Data are presented as mean values. **(B)** Immunofluorescent stainings of NANOG, SOX17 and DNA (DAPI) from day 0/5 in vitro endoderm differentiated SOX17<sup>Δ5CTCF#8.2</sup> or wild-type cells. **(C)** TPM expression heatmap of genes associated within the overall SOX17-TAD. Expression values in iPSC, and CXCR4<sup>-</sup> EN populations are depicted as mean values of Log(SOX17<sup>Δ5CTCF#8.2</sup>/WT)<sub>2</sub>FC across independent experiments ( $n = 3$ ). (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Landshammer A. and Kretzmer H. contributed to this figure)

CXCR4<sup>-</sup> SOX17<sup>Δ5CTCF#8.2</sup> cells may be transcriptionally delayed and just about to enter definitive endoderm. Further we found genes associated with negative regulation of growth to be enriched in SOX17<sup>Δ5CTCF#8.2</sup> cells when comparing wild type and SOX17<sup>Δ5CTCF#8.2</sup> CXCR4<sup>+</sup> populations (437 genes) (Fig. 18D). Most interestingly, we found genes associated with gastrulation and stem cell differentiation (Fig. 18E) enriched in CXCR4<sup>+</sup> SOX17<sup>Δ5CTCF#8.2</sup> cell populations when compared to its less differentiated CXCR4<sup>-</sup> cell populations, which again highlights a developmental delay of SOX17<sup>Δ5CTCF#8.2</sup> CXCR4<sup>-</sup> cells and a compromised ability to generate proper definitive endoderm.

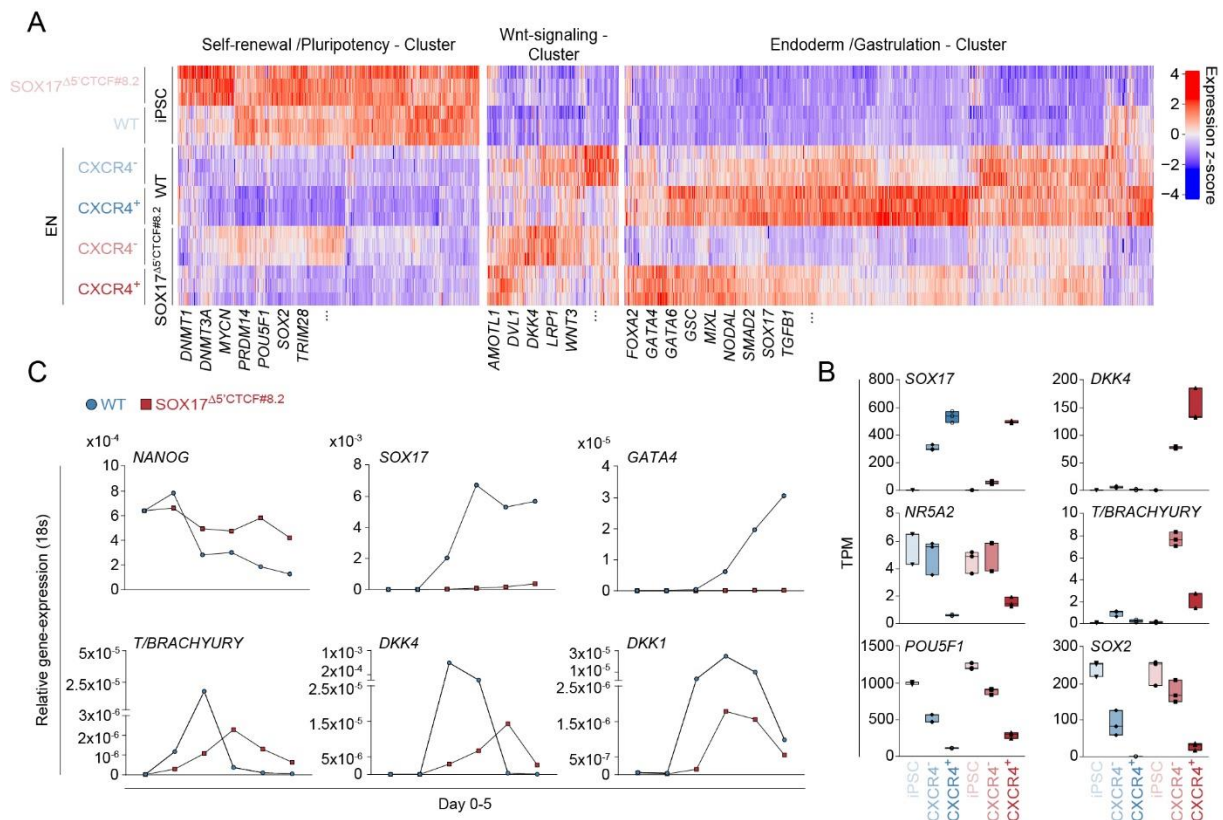
In order to perform gene enrichment analysis in a more unbiased way we performed expression z-score clustering of the most variable genes (4,151 genes) throughout all cell populations (Fig. 19A). We were able to identify three different clusters of genes: Endoderm/Gastrulation (2,282 genes), Wnt-signaling (569 genes) and Self-renewal/Pluripotency (1,300 genes). We found most highly expressed genes within the



**Fig. 18 Comprehensive gene expression analysis of wildtype and Boundary 2 perturbed endoderm. (A)** Principal component analysis of RNA-seq data, depicting sample clusters by the use of the 100 most variable genes. The first two principal components (PCs) are displayed. Arrows and numbers indicate group comparisons. GSEA of differentially expressed genes between compared groups are indicated below; significantly enriched biological processes are depicted in black, pathways in gray. **(B)** TPM Z-score raw normalized hierarchical clustering of the PCA-derived most variable genes (100) across different sub-populations. **(C-E)** GO-terms of GSEA for biological processes from different sub-populations. FDRs are depicted by color. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Landshammer A. and Kretzmer H. contributed to this figure)

endoderm/gastrulation cluster for CXCR4<sup>+</sup> wild type cells while genes for the self-renewal/pluripotency cluster were identified among both iPSC populations to be highest in expression (Fig. 19A). To our surprise we found Wnt-signaling associated genes to be most

highly expressed in both CXCR4<sup>±</sup> SOX17<sup>Δ5<sup>CTCF</sup>#8.2</sup> but also CXCR4<sup>-</sup> wild type populations (Fig. 19A). One of these genes, *DKK4* which is a soluble, canonical Wnt-signaling antagonist, was exclusively upregulated in both SOX17<sup>Δ5<sup>CTCF</sup>#8.2</sup> CXCR4<sup>±</sup> populations (Fig. 19A,B). *DKK4* is known to inhibit the interaction of LRP5/6 with Wnt by forming a ternary complex with the transmembrane protein KREMEN that promotes internalization of LRP5/6[214]. We concluded that increased *DKK4* expression may lead to insufficient canonical Wnt-signaling which is required for proper endoderm differentiation[215]. In concordance with Wnt-signaling, we were able to additionally identify key premature mesendodermal markers, such as *T/BRACHYURY* and *NR5A2* to be highly expressed in SOX17<sup>Δ5<sup>CTCF</sup>#8.2</sup> CXCR4<sup>-</sup> cell populations. This was accompanied by partially decreased but still high levels of pluripotency markers such as *SOX2* and *POU5F1* (Fig. 19A,B). Together with the earlier observations of high pluripotency marker expression and maintenance of an epithelial cell character (Fig. 17A,B) we hypothesized a delayed endoderm differentiation program and Wnt-pathway deregulation.



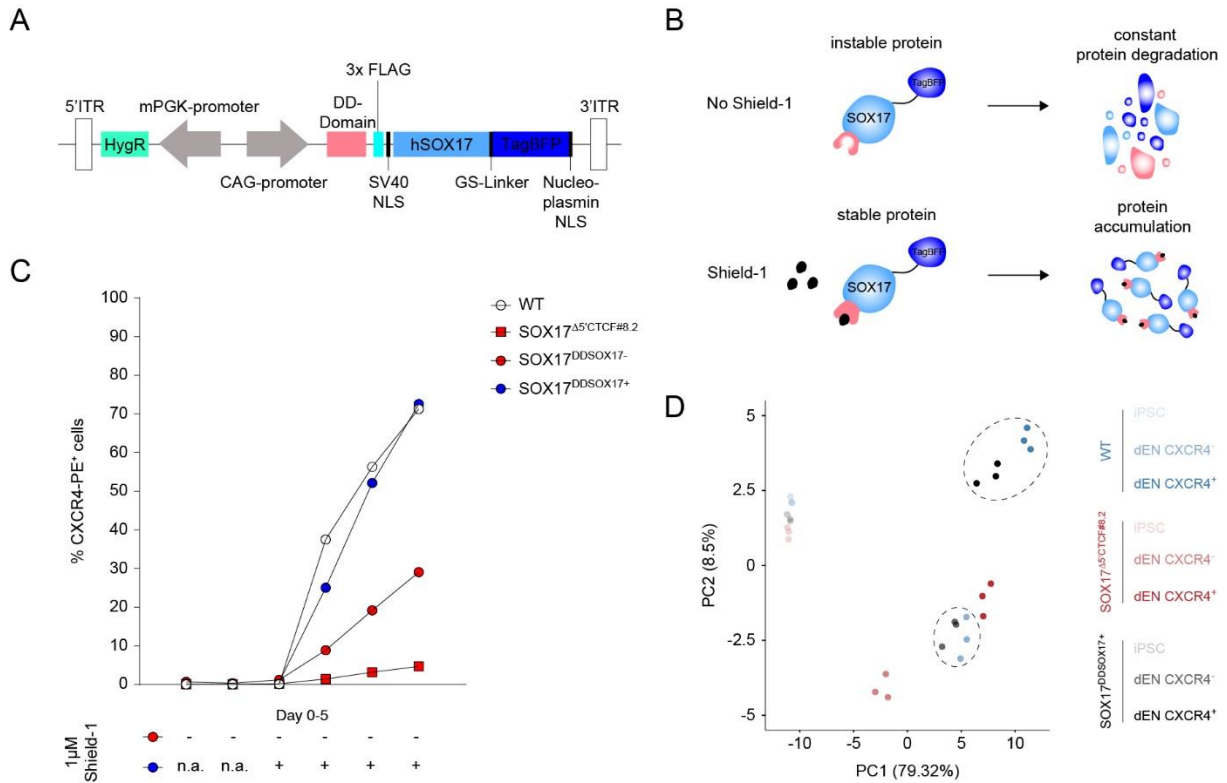
**Fig. 19 SOX17 Boundary 2 caused gene-deregulation leads to a “mesendodermal like” state. (A)** TPM Z-score raw normalized clustering of the most variable genes (4151). Genes were grouped in clusters according to enriched gene-sets; Endoderm cluster (2,282 genes), Wnt signaling cluster (569 genes) and self-renewal /pluripotency cluster (1,300 genes). Gene examples per cluster are depicted on the upper right. **(B)** TPM values shown for a subset of genes ( $n = 3$  biological replicates). The box indicates the interquartile range (IQR), the line inside the box shows the median. **(C)** qRT-PCR of bulk-populations from a subset of genes related to Wnt signaling, mesendoderm, endoderm and pluripotency over 5 days endoderm differentiation. Expression values are depicted as relative gene-expression ( $2^{-\Delta\Delta C_t(\text{GOI}-18s)}$ ) ( $n = 2$  biological replicates)). Data are presented as mean values. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Landshammer A. and Kretzmer H. contributed to this figure)



To test this idea, we performed temporally resolved qRT-PCR analysis of bulk differentiations. Bulk transcriptomes of  $SOX17^{\Delta 5'CTCF\#8.2}$  cells showed a strong expression reduction of the endodermal key markers  $SOX17$  and  $GATA4$  in concordance with stable  $NANOG$  expression over time, compared to wild type (Fig. 19C). Mesendodermal key marker  $T/BRACHYURY$  and Wnt-signaling antagonists  $DKK1/DKK4$  were overall reduced in expression, especially at the onset of differentiation (day 1-3), finally resulting in elevated expression levels at day 5 (Fig. 19C). Taken together these results suggest that  $SOX17^{\Delta 5'CTCF\#8.2}$  cells can partially exit pluripotency but are delayed in their differentiation and trapped in a “mesendoderm-like” state due to Wnt-signaling deregulation. All caused by boundary-perturbation associated mis-expression of  $SOX17$ . In sum we conclude this molecular phenotype to cause definitive endoderm differentiation failure.

To ultimately proof whether the observed phenotype is caused by reduced  $SOX17$  expression levels and may be reversible by restoring endoderm required  $SOX17$  ectopically, we made use of a destabilized ectopic  $SOX17$ -TagBFP expression system. We integrated that system randomly into the  $SOX17^{\Delta 5'CTCF\#8.2}$  genetic background using the Piggy-BAC transposase (Fig. 20A). TagBFP<sup>-</sup> Hygromycin resistant  $SOX17^{\Delta 5'CTCF\#8.2}$  cells, were sorted, cultured in bulk and further referred as  $SOX17^{DDSOX17}$ . Cells differentiating towards endoderm from day 2 onwards were either treated ( $SOX17^{DDSOX17+}$ ) or not treated ( $SOX17^{DDSOX17-}$ ) with a small molecule<sup>[216]</sup> named Shield-1, in order to reverse the constitutive ectopic  $SOX17$ -TagBFP degradation or not, respectively (Fig. 20B). When performing FACS analysis, we observed elevated CXCR4<sup>+</sup> fractions even in untreated  $SOX17^{DDSOX17-}$  cells compared to the original knock-out  $SOX17^{\Delta 5'CTCF\#8.2}$ , which indicated partial leakiness of our expression system (Fig. 20C). However, CXCR4<sup>+</sup> fractions were observed to be restored to almost wild type levels in  $SOX17^{DDSOX17+}$  cells, highlighting functionality of the system and potential rescue of our earlier observed phenotype by ectopic  $SOX17$ -TagBFP (Fig. 20C). To investigate the rescue extent in  $SOX17^{DDSOX17+}$  CXCR4<sup>±</sup> cell fractions on a global transcriptional level, we again performed PCA of the 100 most variable genes across wild-type,  $SOX17^{\Delta 5'CTCF\#8.2}$  and  $SOX17^{DDSOX17+}$  cells. As expected, we found both populations including the undifferentiated iPSCs to cluster closely with their wild type matching cell populations (Fig. 20D).

Altogether, we have shown that  $SOX17$  CTCF loop domain perturbation leads to enhancer-associated deregulation of  $SOX17$  gene-control during the formation of definitive endoderm. Following loss of  $SOX17$  we found reduced potential of iPSCs to exit pluripotency, Wnt-signaling de-regulation, potentially leading to a trapped “mesendodermal-like” state and ultimate definitive endoderm differentiation failure, reversible by ectopic  $SOX17$  expression.



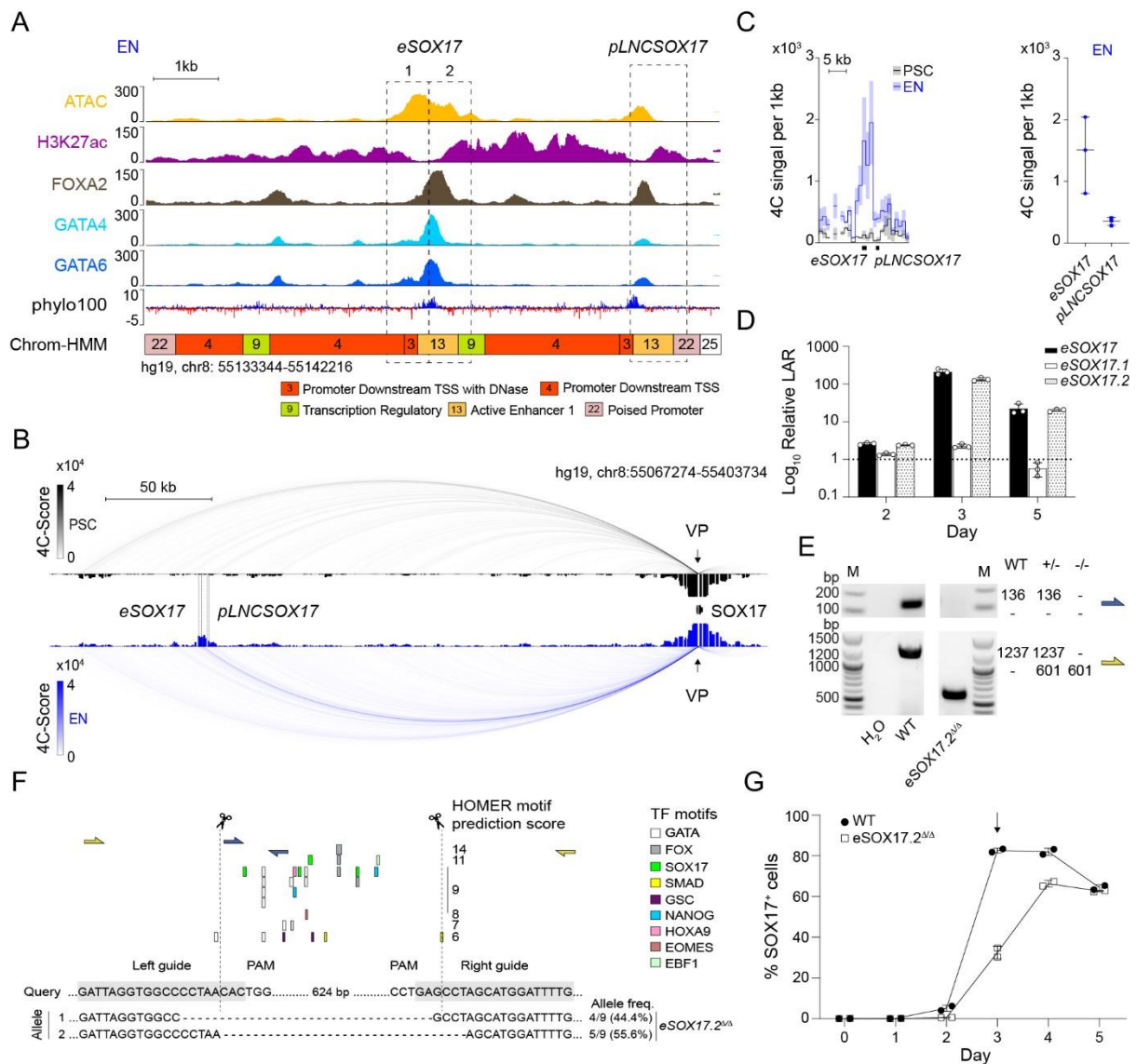
**Fig. 20 SOX17 Boundary 2 perturbation rescue experiment. (A)** Scheme of the SOX17 rescue construct, flanked by PiggyBAC ITRs including a Hygromycin selection cassette. Ectopic SOX17 is a fusion protein made of SOX17, TagBFP fluorescent reporter and a destabilizing domain (DD). Included are a 3x FLAG sequence and a nuclear localization signal (SV40 NLS). **(B)** Model of the SOX17 rescue construct showing constant ectopic SOX17 degradation which is exclusively stabilizable by Shield-1. **(C)** Fluorescence activated cell sorting (FACS) time-course data of wild-type, SOX17 $\Delta 5/CTCF\#8.2$  and Shield-1 treated/untreated SOX17 $^{DDSOX17}$  iPSCs during directed differentiation towards definitive endoderm. CXCR4 (CD184) is depicted as percentage CXCR4 $^{+}$  in bulk cell populations. Symbols represent the mean across independent experiments ( $n = 2$ ) **(D)** Principal component analysis of RNA-seq data, depicting sample clusters by the use of the 100 most variable genes. The first two principal components (PCs) are displayed. Dashed circles highlight clustering proximity between SOX17 $^{DDSOX17+}$  and wild-type cells compared to corresponding SOX17 $\Delta 5/CTCF$  in definitive endoderm. (Adapted figure from Wu H.J. and Landshammer A. et al., 2021. Landshammer A. and Kretzmer H. contributed to this figure)

#### 4.4 Epigenetic profiling of the *SOX17*-DMR and identification of a novel lncRNA locus

As outlined in the last chapters our previous investigations led to the identification of a definitive endoderm (EN) specific differentially methylated region (DMR) located 230 kb upstream of the *SOX17* gene (*SOX17*-DMR) (Fig. 9), within a topological isolated CTCF loop-domain insulated by strong CTCF-boundaries[29, 91] (Fig. 15C). Previous efforts investigating DNA methylation changes during human three germ-layer differentiation, pinpointed distinct changes occurring at germ-layer specific regions distributed through the genome (Fig. 8 and Fig. 9).[29] *SOX17*-DMR has been associated with the decoration of EN-specific H3K27ac histone modifications (DRE A) and physical interaction/proximity with the *SOX17* promoter (*pSOX17*), indicating its regulatory function as a distal *SOX17* enhancer [29, 91] (Fig. 15C). Additional investigations demonstrated that CTCF loop-domain perturbation concomitant with DRE-*pSOX17* interaction loss, leads to *SOX17* gene deregulation and severe EN differentiation failure[91]. To gain deeper insights and to assess *SOX17*-DMR's functional relevance, we further in depth characterized and studied *SOX17*-DMR in an intact, wild type CTCF loop domain setting.

In particular, the *SOX17*-DMR was found to be 6,2 kb long including 55 CpGs[29] (Fig. 8 and Fig. 9). The region is characterized by EN-specific DNA methylation loss which we found accompanied by epigenetic remodeling to a transcriptionally active state<sup>1</sup> (Fig. 21A). On top, we found the *SOX17*-DMR to be occupied by many endoderm specific TFs e.g., FOXA2, GATA4, GATA6 (Fig. 21A) but also EOMES, SMAD2/3 and SMAD4 (not shown). Interestingly, these TFs have previously been shown to deploy a transcriptional network governing cardinal endodermal genes such as *SOX17*. [179, 217] Inside the *SOX17*-DMR, we identified two very distinct sites of high TF-occupancy and open chromatin, which show enriched vertebrate sequence conservation but two different chromatin-states[218·219] (Fig. 21A).

The first site further referred as enhancer-DMR (*eSOX17*), is characterized by the presence of a strong enhancer signature with open chromatin (ATAC-seq signal) flanked by H3K27ac (Fig. 21A) and H3K4me3 (Fig. 22A) but without any presence of H3K4me1 (not shown), supported by the respective Chrom-HMM state (Fig. 22A). To elucidate whether this region can physically interact with *pSOX17*, hence having the potential to act as an enhancer, we performed highly resolved circular chromatin conformation capture (4C) sequencing in the female iPSC line ZIP13K2[211] further referred as PSCs and PSC-derived EN cells (Fig. 21B). Utilizing the *SOX17* promoter as 4C-viewpoint (VP), we found a specific interaction between the *eSOX17* and the *pSOX17* in an endoderm specific fashion (Fig. 21B,C). To functionally prove, *eSOX17*'s enhancer potential, we divided the region into two parts (*eSOX17.1/eSOX17.2*) based on TF occupancy and tested their enhancer-based relative luminescence activity ratio (LAR) in PSC-derived EN cells (Fig. 21D). At day 3 of EN differentiation *eSOX17.2* (rel. LAR



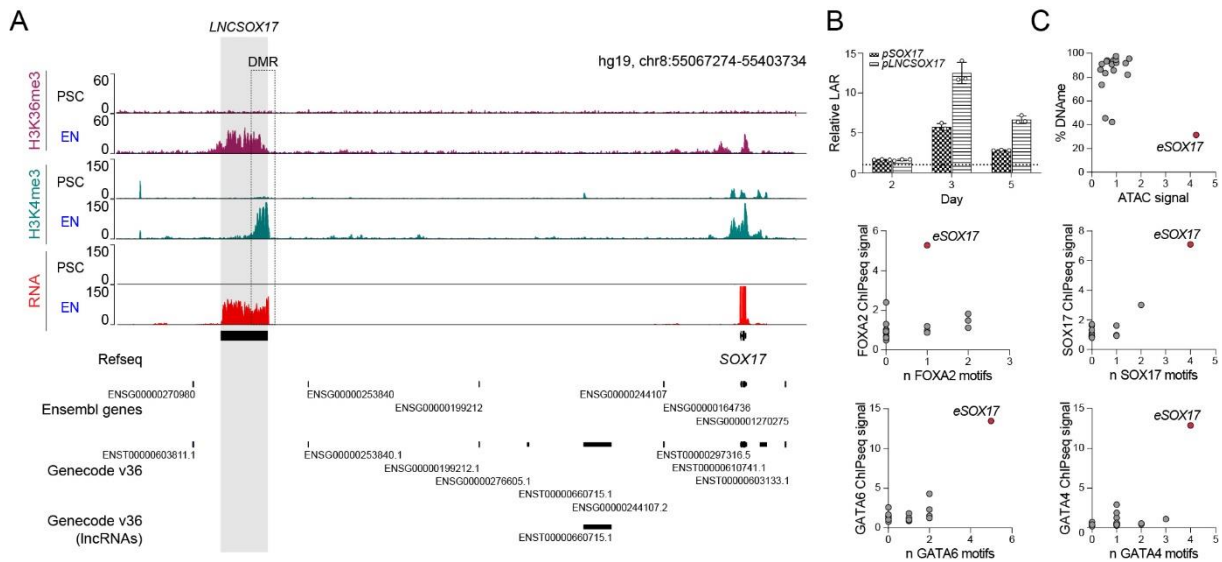
**Fig. 21 Epigenetic profiling and genetic characterization of the SOX17-DMR.** (A) Zoomed in view of the SOX17 DMR in definitive endoderm (EN) comprising ATAC sequencing profile and H3K27ac, FOXA2, GATA4 and GATA6 ChIP sequencing profiles. Chrom-HMM state profile is shown below the phylo100 UCSC conservation track. Dashed lines indicate the two distinct regulatory elements, characterized by enriched transcription factors occupancy (eSOX17 and pLNCsox17). (B) 4C sequencing of PSC (black) and EN (blue) at the SOX17-locus. Normalized interaction-scores displayed as arcs and histogram-profiles utilizing the SOX17 promoter as viewpoint (VP). (C) 4C interactions as a zoomed in view at the SOX17 regulatory element (left) and corresponding quantification (right). Lines represents the median and the shaded areas depict 95% CI; in the quantification (left), the central line represents the median and error bars show SD across independent experiments (right) ( $n = 3$ ). (D) Firefly luciferase assay from either eSOX17.1/2 or both together at day 2, 3 or 5 of EN differentiation. Values are calculated as luciferase activity ratio (LAR) between Firefly and Renilla signal, finally normalized on the empty vector background signal. Dashed line shows empty vector signal at “1”. Bars indicate mean values, error bars show standard deviation (SD) across independent experiments ( $n = 3$ ). (E) Genotyping generated by two different primer-pairs to profile eSOX17.2 genetic ablation. (F) Schematic of the CRISPR/Cas9 based eSOX17.2 perturbation strategy. sgRNA sequences are highlighted in grey while Cas9 targeting sites are depicted by dashed lines. Sanger sequencing results and detected allele-frequency are summarized below. Homer TF motif prediction score numbers are given, and TFs indicated by respective colors. (G) SOX17 FACS analysis of EN differentiating wild-type and eSOX17.2<sup>Δ/Δ</sup> cell fractions, median and error bars (SD) are show across independent experiments ( $n = 2$ ). (Landshammer A., Bolondi A., Wu H.J., Parsi K.M., Huey J. contributed to this figure)

150,49) but not *eSOX17.1* (rel. LAR 2,37) showed elevated enhancer activity (Fig. 21D). The enhancer activity decreased again within day 5 of EN differentiation to a similar extend for *eSOX17.2* (rel. LAR 17,55) and the entire *eSOX17* (rel. LAR 21,38) (Fig. 21D). Moreover, a homozygous deletion of *eSOX17.2* (Fig. 21E,F) led to a drastic reduction of *SOX17*<sup>+</sup> cell populations (-51,80 % delta) at the onset of *SOX17*-expression in our temporally resolved *in vitro* EN differentiation compared to wild type PSC-derived EN cells (Fig. 21G). Taken together our data confirms earlier findings of *eSOX17* to be an actual, fully functional, and transiently active developmental distal enhancer, necessary for the early activation of *SOX17* during the formation of *in vitro* EN.

The second region, further referred as *pLNC**SOX17*, downstream of the *eSOX17* indicates the presence of a promoter signature with lower levels of open chromatin (ATAC-seq signal), H3K27ac (Fig. 21A) and H3K4me3 (Fig. 22A) but instead to *eSOX17* with high signal for H3K4me1 (not shown), supported by the respective Chrom-HMM state (Fig. 21A). Intrigued by the promoter signature we sought whether we would also find a chromatin signature for gene-bodies. To our surprise we were able to find high levels of H3K36me3 spanning, both regions comprising a total region of 22 kb (Fig. 22A). To find out whether these 22 kb could potentially express RNA-transcript we performed poly(A)<sup>+</sup> RNA sequencing on human PSCs and PSC-derived EN cells (Fig. 22A). To our most surprise, we were able to identify a novel 22 kb long RNA-element specifically expressed in EN compared to undifferentiated PSCs (Fig. 22A) whose locus we named *LNC**SOX17*.

To confirm our finding, we tested the promoter activity of the initially identified *eSOX17* downstream region, *pLNC**SOX17* in PSC-derived EN cells (Fig. 21A). We observed strikingly increased promoter activity from *pLNC**SOX17* (rel. LAR 12,79) in endodermal cells at day 3 compared to *pSOX17* (rel. LAR 5,78) (Fig. 22B). Relative LAR levels for both regions decreased to a similar extend at day 5 (Fig. 22B). Nevertheless, since both *pLNC**SOX17* and upstream located *eSOX17* are overlapping with the *LNC**SOX17* locus we finally wanted to assure *eSOX17*'s regulatory identity. Hence, we further carried out combined epigenetic and TF binding profiling along the *LNC**SOX17* locus and found *eSOX17* (red circle) and not *pLNC**SOX17* to be the only DNA element bearing enhancer signature (Fig. 22C).

Overall, we were able to show that the endoderm specific *SOX17*-DMR 230 kb upstream of *SOX17* not only consists of a distal developmental enhancer-element for *SOX17* (*eSOX17*), but also a novel RNA promoter region namely *pLNC**SOX17*. We find the transcriptional activities of both regions to correlate strongly with the presence of *SOX17*<sup>+</sup> cell populations in our *in vitro* EN differentiation system. Most surprisingly, we found a previously unknown 22 kb long RNA-element for the first time, specifically expressed upon definitive endoderm formation which we named *LNC**SOX17*.



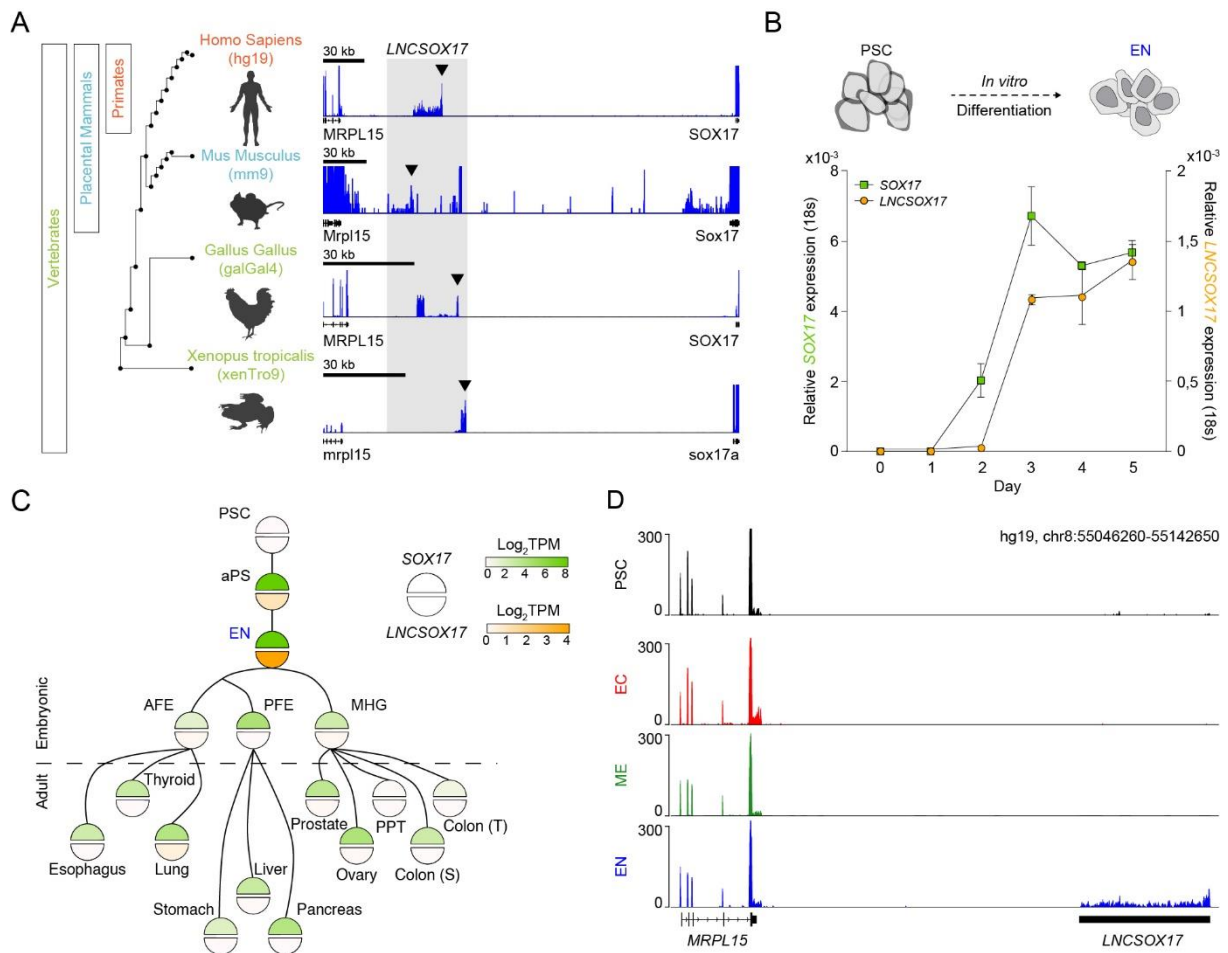
**Fig. 22 Identification of the novel lncRNA locus *LNC5OX17* and functional validation of *pLNC5OX17*.** (A) Epigenetic landscape of the human *SOX17* locus in pluripotent and endoderm cells as depicted by ChIP sequencing tracks of CTCF, H3K36me3 and H3K4me3 as well as RNA sequencing profiles in PSCs and EN. *LNC5OX17* locus is highlighted in grey and the *SOX17*-DMR indicated by dashed lines. (B) Firefly luciferase assay from either *pSOX17* and *pLNC5OX17* at day 2, 3 or 5 of EN differentiation. Values are calculated as luciferase activity ratio (LAR) between Firefly and Renilla signal, finally normalized on the empty vector background signal. Dashed line shows empty vector signal at “1”. Bars indicate mean values and error bars show standard deviation (SD) across independent experiments ( $n = 3$ ). (C) Scatter plots displaying DNA methylation levels, ATAC signal, endoderm TF occupancy as measured by ChIP sequencing and TF binding motifs abundance at the *LNC5OX17* locus in EN cells. The *LNC5OX17* transcribed region was binned into 18 bins (dots) of the same size, including *eSOX17* (red dot). Note how *eSOX17* is depleted of DNA methylation and enriched in ATAC signal, endoderm TF binding motifs and actual TF occupancies as compared to the rest of the transcribed region, indicating a specific enhancer identity. (Landshammer A., Bolondi A., Parsi K.M., Huey J. contributed to this figure)

## 4.5 Characterization of the novel long non-coding RNA (lncRNA) *LNC*SOX17

So far, *SOX17* was known to be the only gene located within the 336 kb *SOX17* loop-domain insulated by strong CTCF-boundaries<sup>70,133</sup> (Fig. 22A). However, closer inspections suggested the presence of another potential gene locus of a 22 kb long transcribed region approximately 230 kb upstream of *SOX17* (Fig. 22A). The epigenetic and transcriptional signature combined with a strong UCSC PhyloCSF sequence conservation pointed to the identification of a previously unannotated, potential intergenic lncRNA (lincRNA), subsequently termed *LNC*SOX17 (Fig. 21A and Fig. 22A). lncRNAs are RNA transcripts  $\geq 200$  nucleotides (nt), they are generally not translated into functional proteins<sup>[158, 159]</sup>, mainly PolIII driven, capped by 7-methyl guanosine (m<sup>7</sup>G) at their 5' ends, bear polyadenylated 3' ends, show generally low expression levels and less sequence conservation compared to coding mRNAs, they are localized mainly nuclear and splicing appears to be less efficient<sup>[163, 164]</sup>.

To further explore the presence of *LNC*SOX17 expression in definitive endoderm of different model organisms, we utilized public transcriptional data of vertebrate stage-matched embryonic tissues, which revealed the presence of a conserved transcript at the locus distal from *SOX17* gene-body close to *MRPL15* (Fig. 23A). We further investigated the expression of the transcript during *in vitro* EN with temporally resolved qRT-PCR and found *LNC*SOX17 expression to be highly EN-specific and following *SOX17* kinetics but with an approximate 24-hour delay (Fig. 23B). Testing for co-occurrence of *LNC*SOX17 and *SOX17*, we compared their expression across a wide range of cell and tissue types ( $n = 44$ ) (Fig. 23C and Fig. 24A). *LNC*SOX17 expression appears tightly restricted to human EN and uncoupled from the much broader expression of *SOX17* in many other endoderm-derived tissues, compared to other endoderm lncRNA-TF couples<sup>[220, 221]</sup> (Fig. 23C and Fig. 24A). Moreover, we utilized public RNA sequencing data from the three pluripotent stem cell derived germ-layers to show that *LNC*SOX17 is not expressed during ectoderm and mesoderm formation, confirming its endoderm specificity during exit from pluripotency (Fig. 23D). scRNA sequencing data from the early human gastrulating embryo <sup>[17]</sup> further confirms *LNC*SOX17's tissue specificity *in vivo* (Fig. 24B upper panel) and read-alignments highlight a certain pattern of splicing (Fig. 24B lower panel).

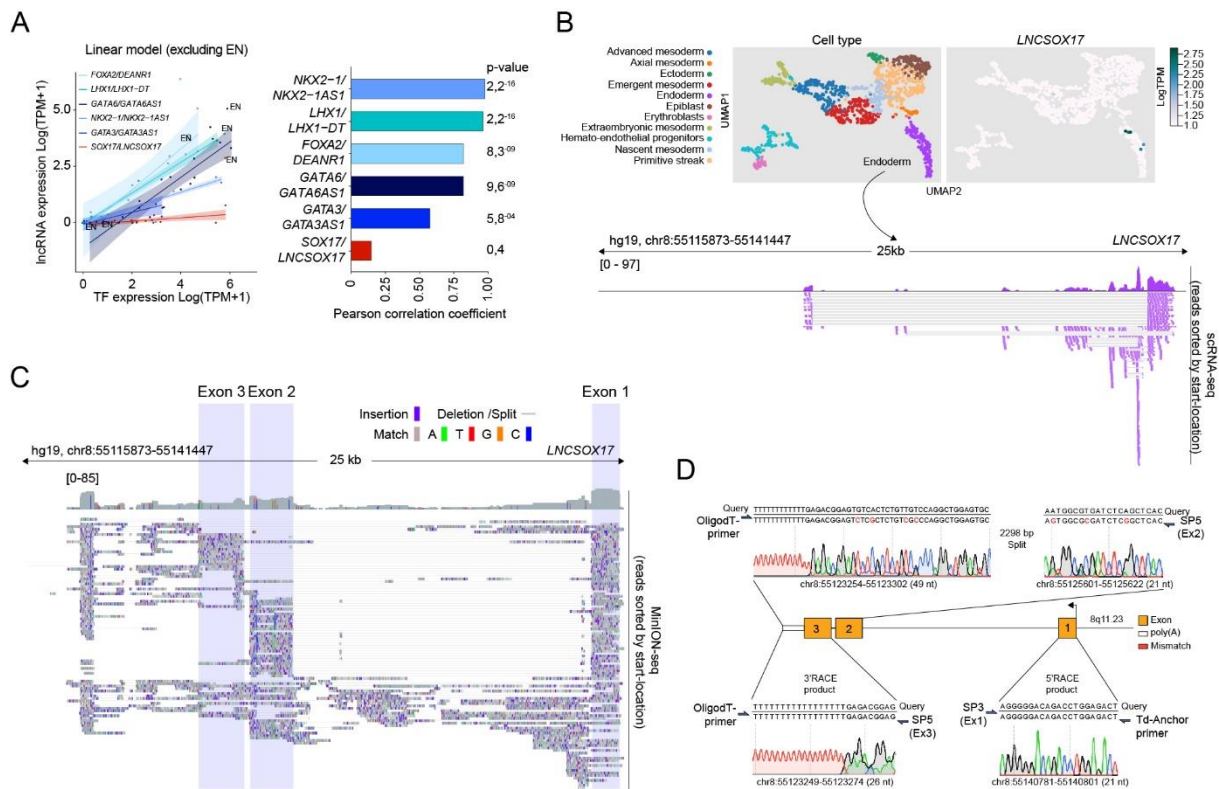
Intrigued by these features, we sought *LNC*SOX17's long-non-coding RNA nature. To do so, we first explored the structure and start/end of *LNC*SOX17 splicing-variants utilizing long-read sequencing of EN cDNA (Oxford Nanopore Technologies), including additional 5'/3' rapid amplification of cDNA end (RACE) PCR followed by molecular cloning and sanger sequencing, respectively. In comparison to the early human embryo data (Fig. 24B lower panel), we found two *bona fide* isoforms by long-read sequencing that account for 23.3% of the split-reads, while



**Fig. 23 *LNCISOX17* shows high definitive endoderm specificity and developmental conservation. (A)** Identification of an unannotated transcript at the *SOX17* locus in embryonic stage RNA sequencing datasets across vertebrates (human (EN), mouse (EN), chicken (HH4) and frog (XT11)). Note how the relative position of the non-coding element (between *SOX17* and *MRPL15* genes) is conserved in all analyzed species. **(B)** Time resolved qRT-PCR profiling *SOX17* (green) and *LNCISOX17* (orange) transcript levels during endoderm differentiation (normalized to the housekeeping gene 18S). Symbols indicate the mean and error bars indicate SD across three independent experiments ( $n = 3$ ). **(C)** Lineage tree heatmap showing *SOX17* (green) and *LNCISOX17* (orange) expression across EN derived embryonic and adult tissues as measured by RNA sequencing derived from a curated data set of the Roadmap Epigenome Project (TPM = transcripts per million). aPS, anterior primitive streak; AFE, anterior foregut endoderm; PFE, posterior foregut endoderm; MHG; mid-hindgut; PPT, Peyer's patch tissue; S, sigmoid; T, transverse. **(D)** Genome browser tracks displaying RNA levels at *LNCISOX17* locus in PSCs and the three germ layers. Note *LNCISOX17* expression specificity as compared to *MRPL15*. (Landshammer A., Kretzmer H. and Bolondi A. contributed to this figure)

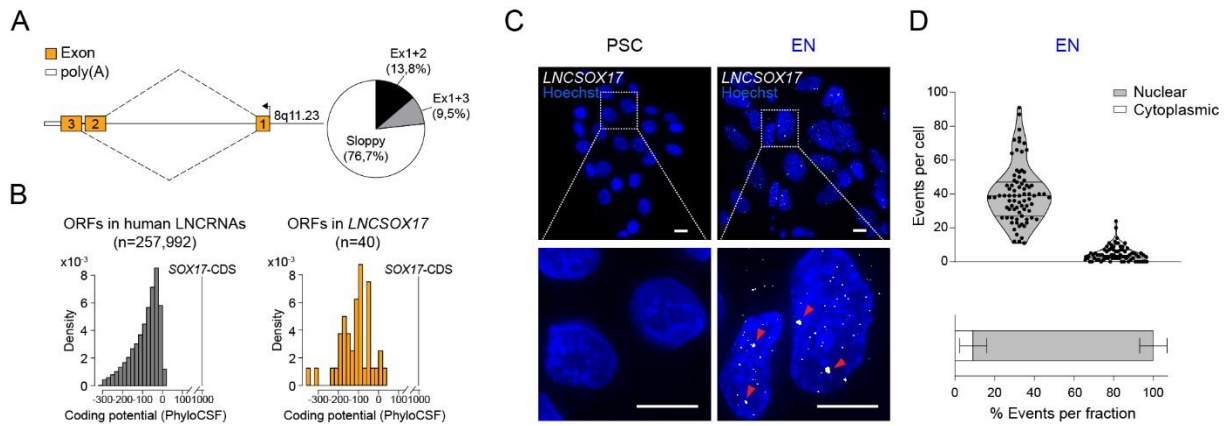
76.7% reads appeared inconsistently spliced; a feature, which is frequently observed in lncRNAs [222-226] (termed "sloppy splicing" Fig. 24C and Fig. 25A). Start and ends as well as the corresponding polyadenylation signal of the two most prevalent isoforms were confirmed by 5'/3' RACE-PCR (Fig. 24D). Next, we assessed the coding potential of *LNCISOX17*. Therefore, we used PhyloCSF and found that 37 of 40 predicted *LNCISOX17* open reading frames (ORFs) would likely result in no functional protein, similarly to other short ORFs (sORFs) in the human lncRNA catalog (Fig. 25B) [227]. Furthermore, even the coding potential





**Fig. 24** *LNC5OX17* is a *SOX17* uncoupled RNA with partially defined processing, start and end. **(A)** The left panel shows a scatter plot with the expression of a set of endoderm lncRNAs (*DEANR1*, *LHX1-DT*, *GATA6AS1*, *NKX2-1AS1*, *GATA3AS1*, *LNC5OX17*) and the corresponding TFs (*FOXA2*, *LHX1*, *GATA6*, *NKX2-1*, *GATA3*, *SOX17*) in the same set of EN tissues of Fig. 22C. In the right panel it is shown a linear model excluding the expression in EN fit for each lncRNA-TF couple. Pearson correlation coefficients as well as corresponding *p*-values are displayed in the bar-plot. Note that the *LNC5OX17*-*SOX17* couple has the lowest degree of tissue co-expression. **(B)** Uniform Manifold Approximation and Projection for Dimension Reduction (UMAPs) showing cell states (upper left panel) and *LNC5OX17* expression (upper right panel) in cells derived from a human gastrulating embryo. Single-cell RNA sequencing track from cells belonging to the endoderm cluster showing reads mapping to the *LNC5OX17* locus (bottom panel). **(C)** MinION long-read sequencing read-track showing *LNC5OX17* coverage and structure in endodermal cells. Sequencing read distribution histogram (top) and individual reads sorted by their start-location (bottom) are displayed. Exon 1, 2 and 3 are highlighted by shading boxes. Sequence mismatches and matches are color coded as described. Split-reads and deletions are shown as thin horizontal lines. **(D)** Sanger sequencing of 3'/5' RACE PCR products. Amplicon specific sequencing results are shown below the query sequence (hg19). Sequencing mismatches are highlighted in red. Primer pairs relative positions used for the PCRs are shown for each product. Sanger sequencing chromatogram color code is used to show the raw reads data. (Landshammer A., Kretzmer H., Tornisiello R., Braendl B., Giesselmann P. and Bolondi A. contributed to this figure)

of the remaining three sORFs is about two orders of magnitude lower than for the *SOX17* coding sequence (Fig. 25B). To investigate cellular localization of *LNC5OX17*, we carried out single-molecule RNA fluorescence *in situ* hybridization (smRNA-FISH) and found it highly enriched at foci within the nuclear compartment, a characteristic feature of non-coding transcripts (median of 40 foci/cell, Fig. 25C,D).



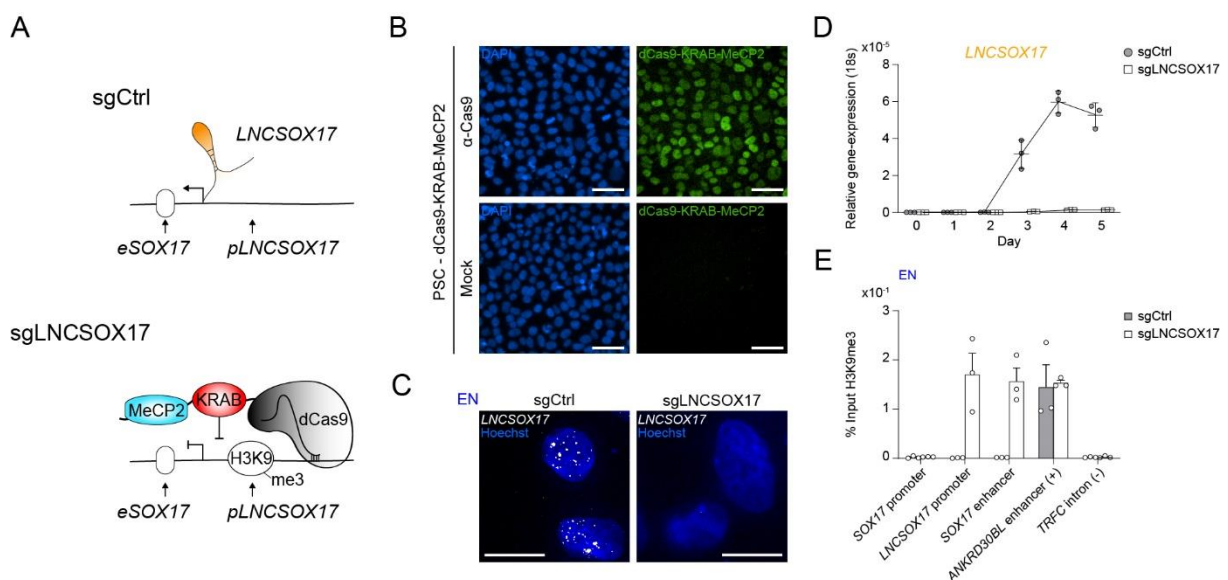
**Fig. 25 *LNC5OX17* is a nuclear long non-coding RNA.** (A) Schematic of *LNC5OX17* isoform structure constructed from MinIONseq reads of endoderm cDNA. Exons are shown in orange while poly(A) is shown in white. The arrow indicates the transcriptional start site (TSS). Pie chart shows isoform reads (Ex1+2 black  $n = 16$ , Ex1+3 grey  $n = 11$ ) and “sloppy spliced” (white  $n = 89$ ) transcript distribution as measured by MinIONseq. (B) Bar plots showing coding potential scores of randomly sampled *LNC5OX17* ORFs ( $n = 257,992$ ) (grey) versus *LNC5OX17* ORFs ( $n = 40$ ) (orange). Scores are shown on the x-axis while ORF-density is plotted on the y-axis. Both conditions area is equal and compared to *SOX17* ORFs as coding gene control. (C) smRNA-FISH of *LNC5OX17* in PSCs (left) and EN cells (right) counter-stained with Hoechst. Red arrowheads indicate two brighter and bigger foci present in each cell, potentially representing sites of nascent transcription. Scale bars indicate  $10\mu\text{m}$ . (D) Frequencies of *LNC5OX17* smRNA-FISH foci in the nuclear (grey) or the cytoplasmic (white) compartments ( $n = 79$  analyzed cells). Lines of the violin plot indicate interquartile range around the median value. In the stacked barplot, error bars indicate SD around the mean value. (Landshammer A., Braendl B., Giesselmann P., Mackowiak S., Much C. and Bolondi A. contributed to this figure)

In summary, our results suggest that the 22 kb *LNC5OX17* locus produces a nuclear long intergenic non-coding RNA (lincRNA), of two highly processed EN-specific, 5' m<sup>7</sup>G-capped and 3' polyadenylated RNA isoforms, including various “sloppily spliced” transcripts.

#### 4.6 *LNC*SOX17 does not regulate *SOX*17 in *cis* during definitive endoderm

To investigate the functional role of *LNC*SOX17 during EN formation, we generated a cell line carrying a constitutive transcriptional repressor (dCas9-KRAB-MeCP2)[228] and then derived two clonal cell lines from it, harboring either a sgRNA targeting a control (sgCtrl), derived from a randomization approach of human TSS regions[229], or the *LNC*SOX17 promoter (sgLNC*SOX*17) (Fig. 26A). Immunofluorescent staining for dCas9 demonstrated its homogeneous expression in the parental cell line (Fig. 26B). The dCas9 mediated silencing resulted in a strong repression of *LNC*SOX17 RNA compared to the control, which was further validated by smRNA-FISH (Fig. 26C,D). Functional validation of our repression system revealed H3K9me3 enrichment around *pLNC*SOX17 in sgLNC*SOX*17 cells, with a certain degree of spreading towards the *eSOX*17 but no apparent consequence on the *SOX*17 regulation (Fig. 26E and Fig. 27D).

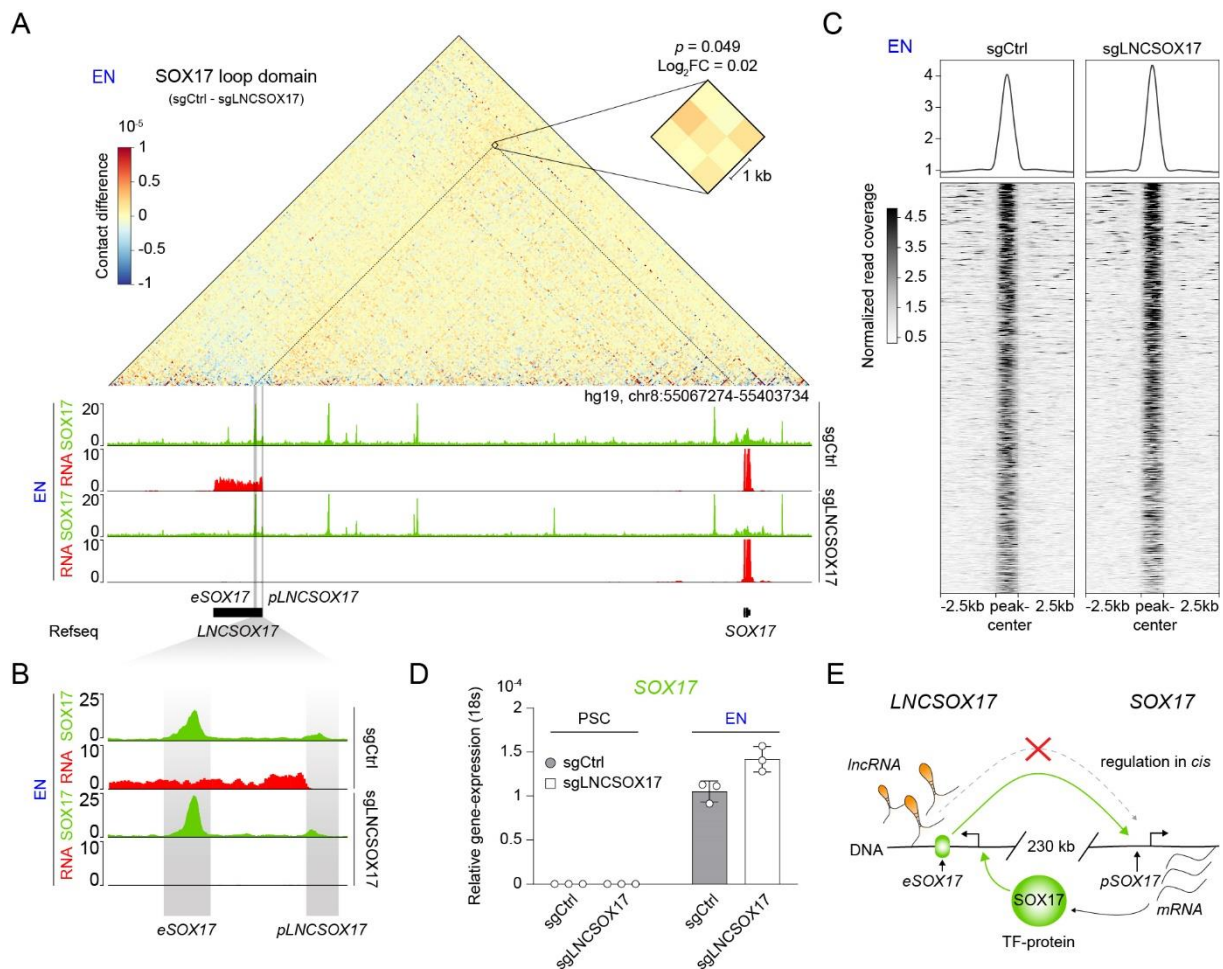
To assess possible effects on *SOX*17 gene-control, we performed Capture Hi-C (cHi-C) in both cell lines. We could not observe any significant interaction differences ( $\text{Log}_2\text{FC} = 0.02$   $p = 0.049$ ) between the two cell lines within the *SOX*17 CTCF loop domain in definitive endoderm



**Fig. 26 CRISPR interference (CRISPRi) based repression of *LNC*SOX17.** (A) Schematic of *LNC*SOX17 locus regulation in the absence (top) or presence (bottom) of a targeting dCas9-KRAB-MeCP2 complex, decorating the *LNC*SOX17 promoter with an H3K9me3 mark 355 bp upstream of the TSS. (B) IF-staining for dCas9 in PSCs expressing dCas9-KRAB-MeCP2 and counter-stained with DAPI. Mock samples represent secondary antibody only controls. Scale bars indicate 50 μm. (C) smRNA-FISH of *LNC*SOX17 in sgCtrl (left) and sgLNC*SOX*17 (right) EN cells counter-stained with Hoechst. Scale bars indicate 10 μm. (D) Time-resolved qRT-PCR showing the expression of *LNC*SOX17 during EN differentiation in the presence or absence of dCas9-KRAB-MeCP2 complex targeting *pLNC*SOX17 (normalized to the housekeeping gene 18S). Symbols indicate the mean and error bars indicate SD across independent experiments ( $n = 3$ ). (E) H3K9me3 ChIP-qPCR enrichment percentages over input is represented at different regions of the genome in sgCtrl and sgLNC*SOX*17 endoderm cells. Bars indicate mean values; error bars indicate the SD across independent experiments ( $n = 3$ ). (Landshammer A. and Bolondi A. contributed to this figure)

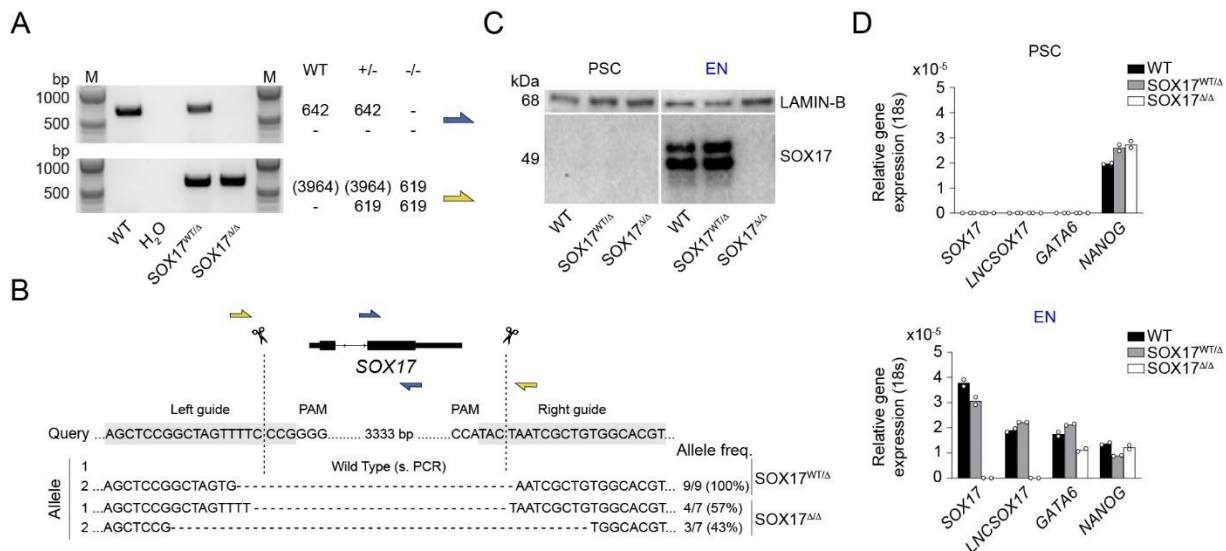
(Fig. 27A). Expression analysis of *LNCXOX17* revealed no effect on *SOX17* expression levels, indicating preserved enhancer functionality and gene-regulation in *cis* even due to loss of *LNCXOX17* expression (Fig. 26D and Fig. 27D).

Next, we performed *SOX17* Chromatin Immunoprecipitation (ChIP) sequencing of our control and loss of function cell line. This demonstrated that there is no change in *SOX17* occupancy at the *SOX17* locus as well as genome-wide occupancy despite the loss of *LNCXOX17* (Fig.



**Fig. 27 CRISPRi repression of *LNCXOX17* does not influence *SOX17* gene control in *cis*.** (A) The upper panel shows a chi-C sequencing subtraction map of the EN sgCtrl-sg*LNCXOX17* at the *SOX17* loop domain. *eSOX17* loop interaction with *SOX17* promoter is shown in the magnification and highlighted by the dotted lines (significance threshold:  $\log_2FC \pm 0.5$ ,  $p < 0.01$ ). In the lower panel *SOX17* EN ChIP sequencing (RPKM) and RNA sequencing (CPM) profiles in the two conditions are shown in the tracks. *eSOX17* and *pLNCXOX17* are highlighted in grey. (B) *SOX17* ChIP sequencing and RNA sequencing tracks at the zoomed in *LNCXOX17* locus showing *SOX17* binding at the *SOX17* enhancer (*eSOX17*) and *LNCXOX17* promoter (*pLNCXOX17*). *SOX17* binding on *pLNCXOX17* results in *LNCXOX17* activation, if *pLNCXOX17* is not targeted by dCas9-KRAB-MeCP2. (C) Heatmap showing *SOX17* binding distribution genome-wide in sgCtrl and sg*LNCXOX17* EN. The displayed peaks represent the union of the identified peaks in the two conditions ( $n = 61694$ ). (D) qRT-PCR showing RNA expression of *SOX17* in PSCs and EN cells in the presence or absence of dCas9-KRAB-MeCP2 complex targeting *LNCXOX17* promoter (normalized to the housekeeping gene 18S). Symbols indicate the mean and error bars indicate SD across independent experiments ( $n = 3$ ). (E) Schematic of the potential *cis*-regulation at the *SOX17* locus. Note, there is no potential *cis*-regulation by *LNCXOX17*. (Landshammer A., Kretzmer H. and Bolondi A. contributed to this figure)

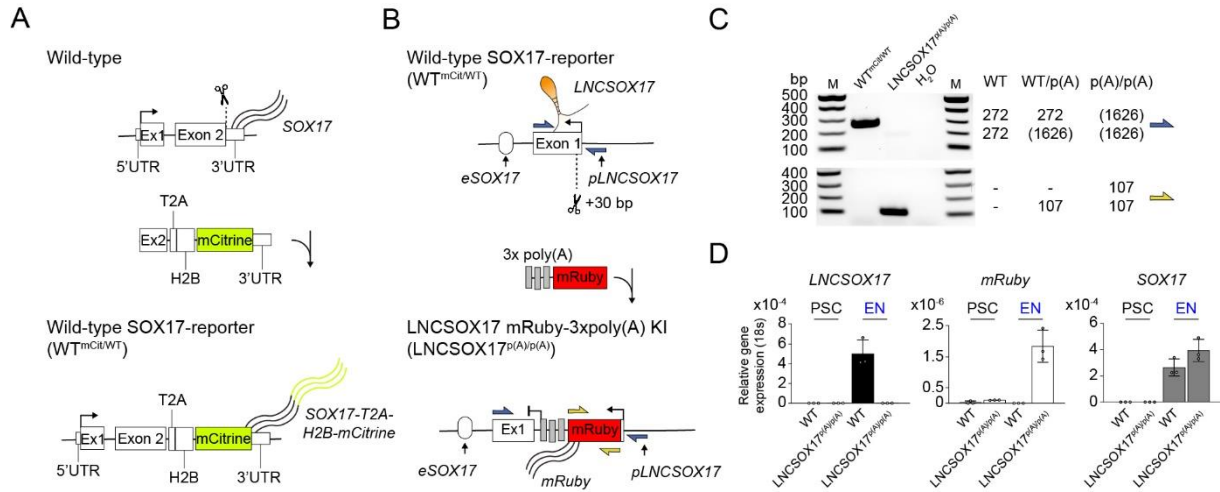
27A-C). Interestingly, we found *SOX17* enrichment at *pLNC*SOX17**, potentially contributing to its activation – consistent with the timed *LNC*SOX17** activation relative to *SOX17* (Fig. 27B and Fig. 23B). To further explore the regulation of *LNC*SOX17** by *SOX17*, we performed sgRNA/Cas9 mediated *SOX17* gene ablation, retrieving heterozygous (*SOX17<sup>WT/Δ</sup>*) and homozygous (*SOX17<sup>Δ/Δ</sup>*) knock out cell lines (Fig. 28A,B). Notably, *SOX17* disrupted cells fail to induce *LNC*SOX17** expression, although EN master TF *GATA6* – upstream of *SOX17* – is activated due to directed differentiation conditions (Fig. 28C,D).



**Fig. 28 CRISPR/Cas9 *SOX17* perturbation shows *SOX17* dependence for *LNC*SOX17** expression.** (A) Genotyping PCR-products, generated by two different primer-pairs to profile *SOX17* gene ablation. Expected amplicon-sizes within a particular genetic background are shown on the side of the agarose gel-picture. (B) Schematic of the Cas9 based *SOX17* gene ablation strategy. sgRNA sequences are highlighted in grey while Cas9 targeting sites are depicted by dashed lines. Sanger sequencing results are summarized below the query-sequence and detected allele-frequency are displayed on the side for each respective genotype. (C) Western Blot showing *SOX17* levels in PSCs and EN cells for the three indicated genotypes. LAMIN-B is used as loading control. (D) qRT-PCR showing *SOX17*, *LNC*SOX17**, *GATA6* and *NANOG* expression in PSCs and EN cells for the three indicated genotypes. Fold change is calculated relative to the 18s housekeeping gene. Bars indicate the means across independent experiments ( $n = 2$ ). (Landshammer A. contributed to this figure)

In order to distinguish between the function of *LNC*SOX17** active transcription and its actual transcript[66, 230], we generated another cell line by introducing a strong transcriptional termination signal downstream of an mRuby cassette in the first exon of *LNC*SOX17**, hereafter *LNC*SOX17*<sup>p(A)/p(A)</sup>* (Fig. 29A-C). This was done in a *SOX17*-T2A-H2B-mCitrine heterozygous genetic wild type background, hereafter *WT<sup>mCit/WT</sup>*. qRT-PCR demonstrated that the expression of *LNC*SOX17** is abolished in homozygous *LNC*SOX17*<sup>p(A)/p(A)</sup>* EN cells, while the mRuby cassette is actively transcribed EN-specifically, indicating ongoing transcription at the locus in an endoderm specific manner (Fig. 29D). In line with our CRISPRi repression experiments, *SOX17* levels are not affected in *LNC*SOX17*<sup>p(A)/p(A)</sup>* EN cells (Fig. 29D and Fig. 27D).

This demonstrates that *LNC*SOX17 induction is most likely dependent on SOX17, whereas the *LNC*SOX17 transcript and the act of transcription are dispensable for eSOX17-pSOX17 interaction and SOX17 activation as well as its genome-wide localization. Therefore, we suggest *LNC*SOX17 not to be a *cis*-acting lincRNA (Fig. 27E).

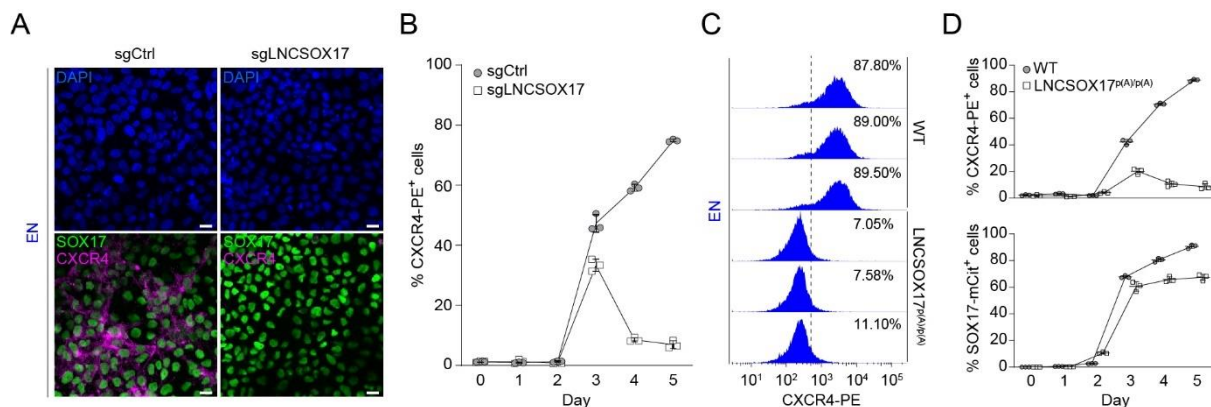


**Fig. 29 CRISPR/Cas9 integrated early transcriptional termination phenocopies repression of *LNC*SOX17.**

**(A)** Schematic of the targeting strategy to generate the SOX17-mCitrine reporter cell line. **(B)** Schematic of the targeting strategy to generate the *LNC*SOX17p(A)/p(A) cell line. **(C)** Genotyping PCR-products, generated by two different primer-pairs to profile the early poly(A) knock-in. Expected amplicon-sizes within a particular genetic background are shown on the side of the agarose gel-picture. **(D)** qRT-PCR showing *LNC*SOX17, *mRuby* and *SOX17* expression in PSCs and EN cells for the two indicated genotypes. Fold change is calculated relative to the 18s housekeeping gene. Bars indicate the means, error bars represent SD across independent experiments ( $n = 3$ ). (Landshammer A. and Bolondi A. contributed to this figure)

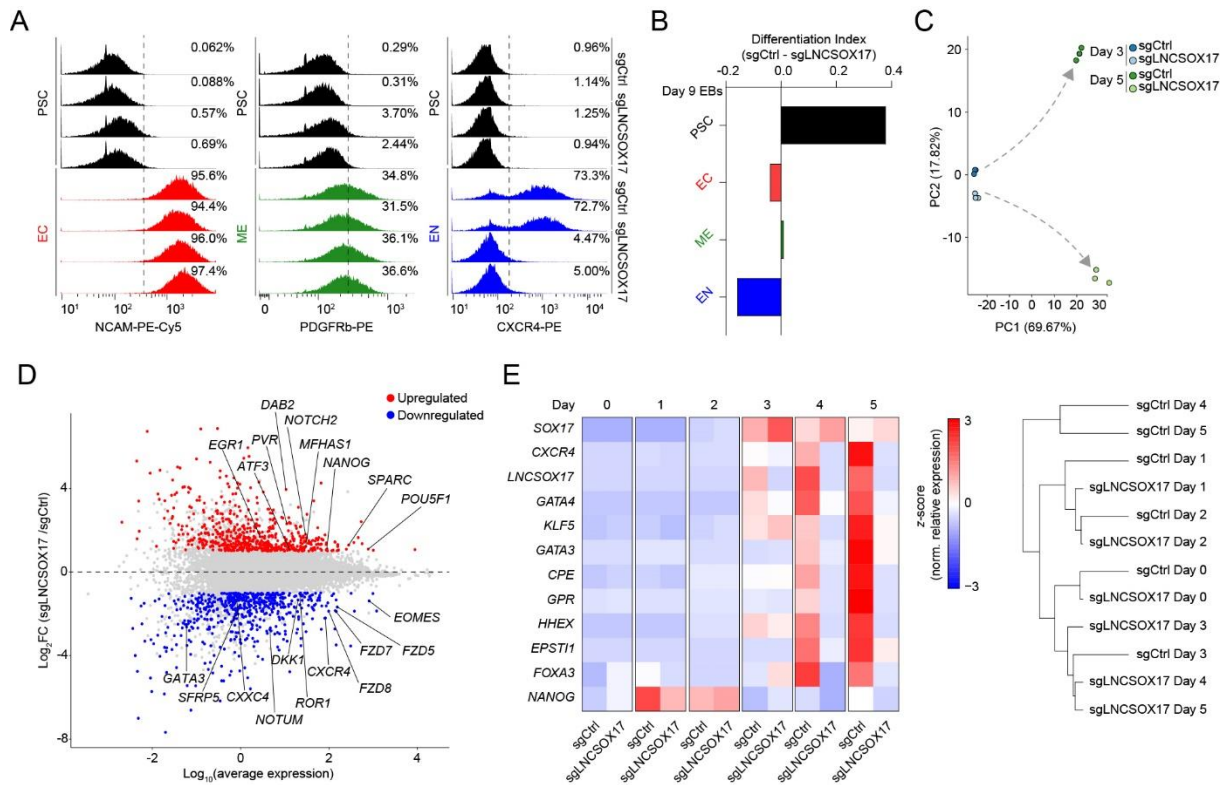
## 4.7 Lack of *LNC*SOX17 leads to aberrant definitive endoderm and differentiation failure

Finding *LNC*SOX17 and its transcription to be dispensable for eSOX17-pSOX17 interaction and the *cis*-regulation of SOX17, we further explored possible functions of *LNC*SOX17. Endodermal lncRNAs as e.g., lncRNA *DIGIT* have been reported to control genes in *trans*[66, 231], hence we carried out immunofluorescent stainings of SOX17/CXCR4 for day 0/5 of EN differentiations and temporally resolved fluorescent activated cell sorting (FACS) for CXCR4, both in control and *LNC*SOX17 depleted cells. The latter showed a substantial reduction in the CXCR4<sup>+</sup> cell population during differentiation, suggesting hampered differentiation potential towards EN (Fig. 30A,B). However, consistent with the transcriptional data (Fig. 27D), SOX17 protein levels were not affected by loss of *LNC*SOX17 (Fig. 30A). Both phenotypes were recapitulated in the *LNC*SOX17<sup>p(A)/p(A)</sup> EN cells (Fig. 30C,D), again confirming the actual RNA and not active transcription at the locus to cause potential endodermal differentiation failure, highlighted by deregulated maintenance of CXCR4 levels.



**Fig. 30 Both, absence, and repression of *LNC*SOX17 lead to CXCR4 deregulation. (A)** Immunofluorescent (IF) staining of SOX17 and CXCR4 in EN cells expressing either sgCtrl or sgLNC SOX17 counter-stained with DAPI. Scale bars, 10 μm. **(B)** Line plot showing percentage of FACS-derived CXCR4<sup>+</sup> cell-population at given time-points during endoderm differentiation (right panel). Symbols indicate mean values, while error bars show SD across independent experiments ( $n = 3$ ). **(C)** FACS histograms showing percentages of CXCR4<sup>+</sup> cells from day 5 EN differentiation of wild-type and *LNC*SOX17<sup>p(A)/p(A)</sup> cell lines. Sample sizes are normalized to 10000 cells /sample. **(D)** FACS time-course experiment showing percentages of CXCR4<sup>+</sup> and SOX17-mCit<sup>+</sup> cells during EN differentiation of wild-type and *LNC*SOX17<sup>p(A)/p(A)</sup> cell lines across independent experiments ( $n = 3$ ). (Landshammer A. contributed to this figure)

As expected, based on its restricted expression, differentiation towards the other two germ-layers (ectoderm and mesoderm) was not affected (Fig. 31A). In contrast, we find little to no change of differentiation propensities for the other germ-layers (ectoderm and mesoderm) to form by *LNC*SOX17 repressed over control cells, utilizing randomly differentiation conditions (Fig. 31B).

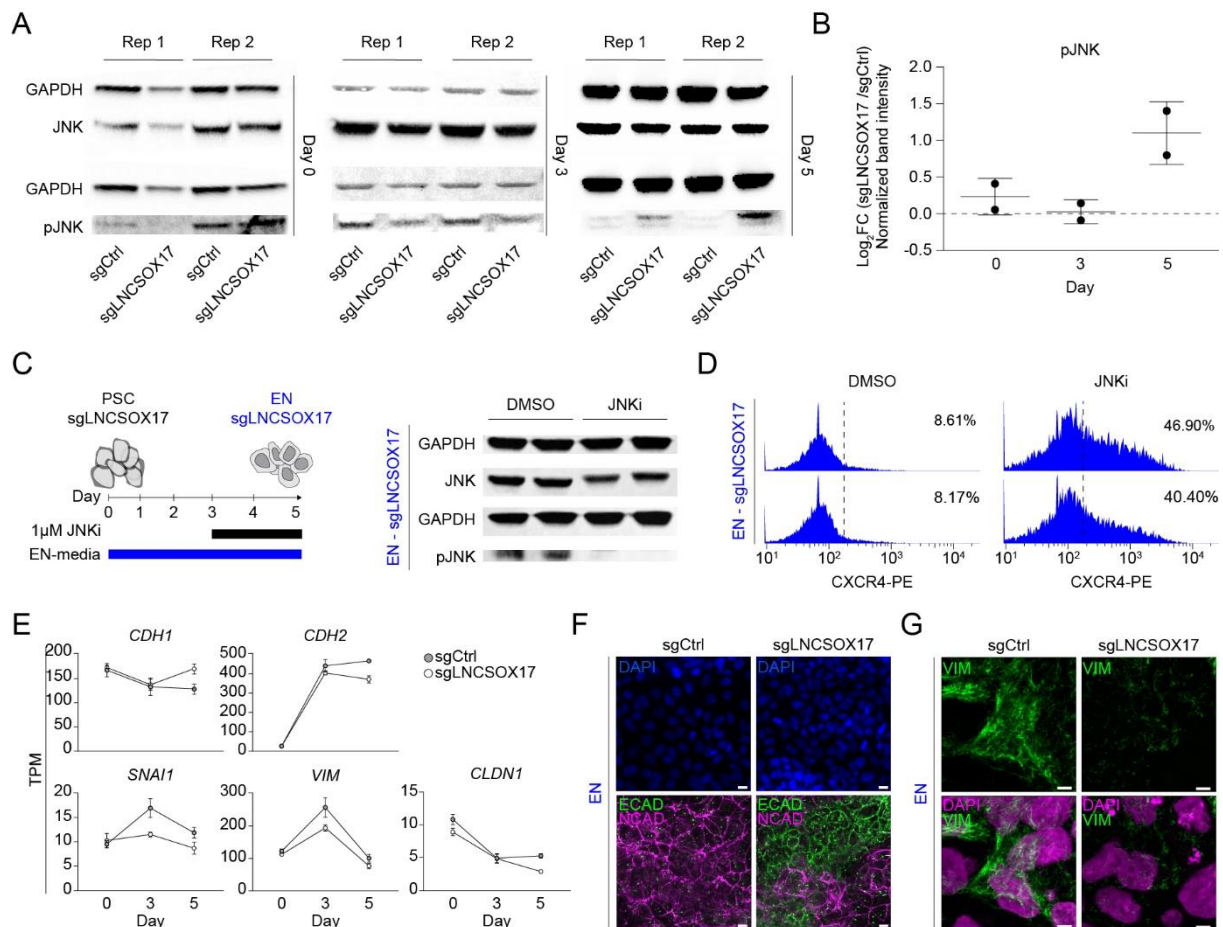


**Fig. 31 Repression of *LNC5OX17* leads to an endoderm specific aberrant transcriptome. (A)** FACS histograms showing percentages of successfully differentiated cells during directed differentiation of sgCtrl and sgLNC5OX17 lines towards the three germ layers. Sample sizes are normalized to 8000 cells /sample. Two independent replicates are displayed. Note how no difference in percentages of differentiated cells is observed for the ectodermal and mesodermal trajectories, while a strong reduction is present in sgLNC5OX17 cells differentiating towards EN. **(B)** ScoreCard assay displaying differentiation index in sgCtrl or sgLNC5OX17 day 9 differentiated embryoid bodies (EBs) ( $n = 48$  EBs per line). **(C)** PCA of the 1000 most variable genes between day 3 and day 5 differentiation of sgCtrl and sgLNC5OX17 as measured by RNA sequencing. Gray dashed arrows indicate the two divergent transcriptomic trajectories. **(D)** Scatter plot highlighting differentially expressed genes between sgLNC5OX17 and sgCtrl EN cells. Significantly ( $\text{Log}_2\text{FC} \geq 1$ ) upregulated genes ( $n = 590$ ) upon *LNC5OX17* repression are shown in red while significantly ( $\text{Log}_2\text{FC} \leq -1$ ) down-regulated genes ( $n = 584$ ) are shown in blue. **(E)** In the right panel, it is shown a heatmap of time-resolved qRT-PCR data for endoderm specific marker genes during EN differentiation of sgCtrl and sgLNC5OX17 cell lines (left) and corresponding hierarchical clustering tree (Euclidean distance) in the tight panel. (Landshammer A., Kretzmer H. and Bolondi A. contributed to this figure)

To characterize the differentiation defect on a molecular level, we performed time-resolved RNA-seq in *LNC5OX17* depleted and control cell lines on day 0, 3, and 5 of endoderm differentiation. Principal Component Analysis (PCA) revealed only marginal variance by day 3, while a more substantial transcriptional divergence was observed on day 5 (Fig. 31C). Analysis of differentially expressed genes (DEGs) identified 584 significantly down- and 590 significantly upregulated genes in *LNC5OX17* depleted cells at day 5 (Fig. 31D). In particular, we found pluripotency genes (e.g., *POU5F1*, *NANOG*) and endoderm/Wnt related genes (e.g., *EOMES*, *GATA3*, *CXCR4*, *FZD5*, *FZD7*, *FZD8*, *DKK1*, *NOTUM*, *ROR1*, *CXXC4*, *SFRP5*) to be significantly up- and downregulated, respectively (Fig. 31D). Time resolved qPCR analysis



over 5 days confirmed, a lack of key endoderm marker activation and expression in *LNC SOX17* depleted cells (including *CXCR4*, *GATA3*, *GATA4*, *KLF5*, *CPE*, *GPR*, *HHEX*, *EPSTI1*, *FOXA3*) [193, 232-237] (Fig. 31E left panel). Further hierarchical clustering revealed day 0-3 samples to cluster according to time while day 4-5 samples to cluster according to sample-type (Fig. 31E right panel) in concordance with our PCA analysis (Fig. 31C).

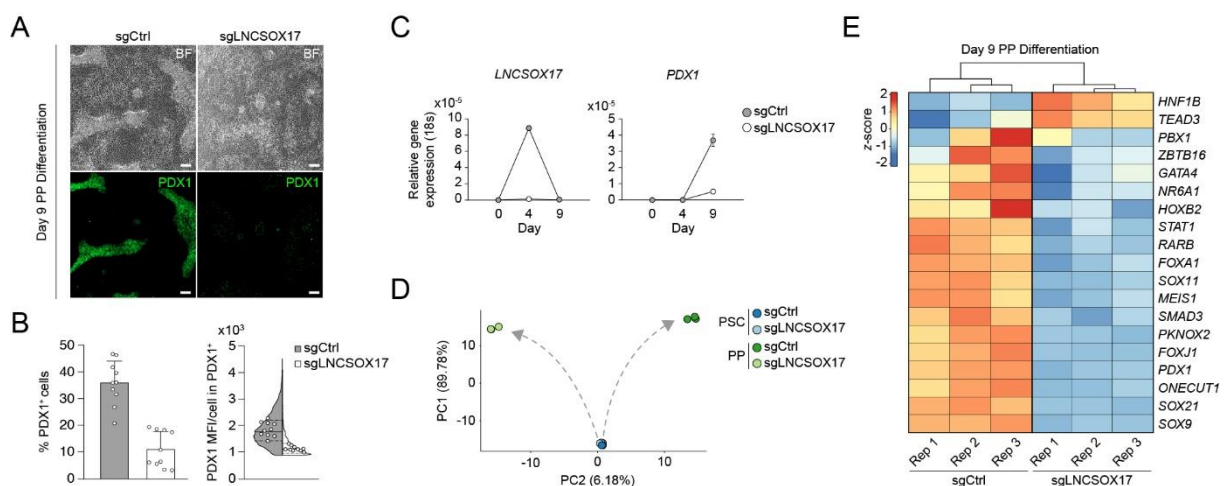


**Fig. 32 Repression of *LNC SOX17* leads JNK hyperactivity and EMT failure. (A)** JNK and pJNK Western Blots of sgCtrl and sgLNC SOX17 day 0,3,5 EN cells. GAPDH signals are used as loading controls above the corresponding JNK/pJNK signals. Independent biological replicates are given per Blot ( $n = 2$ ). **(B)** Boxplot showing relative pJNK level quantification during endoderm differentiation day 5 shown as Log<sub>2</sub>FC of sgLNC SOX17 over sgCtrl. Central line indicates the mean, error bars indicate the SD across independent experiments ( $n = 2$ ). **(C)** Western blot showing the levels of JNK and pJNK during EN differentiation in sgLNC SOX17 cell line in the presence or absence of JNK Inhibitor XVI from day 3 of EN differentiation (see schematic). Independent replicates are displayed ( $n = 2$ ). GAPDH is used as loading control. **(D)** FACS histograms showing percentages of successfully differentiated cells during directed differentiation of sgLNC SOX17 line towards EN in the presence or absence of JNK Inhibitor XVI as measured by CXCR4<sup>+</sup> cells, sample sizes are normalized to 8000 cells /sample. Independent replicates are displayed ( $n = 2$ ). Note how the treatment with JNKi partially rescues the differentiation phenotype in cells lacking *LNC SOX17*. **(E)** Line plots displaying time-resolved expression of selected marker genes in sgCtrl and sgLNC SOX17 cells as measured by RNA-seq (TPM). Marker genes categories are indicated. Symbols indicate mean values and error bars represent SD across three independent experiments. **(F)** IF staining of ECAD and NCAD in EN cells expressing either sgCtrl or sgLNC SOX17 counter-stained with DAPI. Scale bars indicate 10μm. **(G)** IF staining of VIM in EN cells expressing either sgCtrl or sgLNC SOX17 counter-stained with DAPI. Scale bars indicate 5μm. (Landshammer A., Kretzmer H. and Bolondi A. contributed to this figure)

Interestingly, among the significantly, upregulated genes in *LNC SOX17* depleted cells, we found an enrichment of *JUN* (AP-1) pathway target genes (including *EGR1*, *ATF3*, *PVR*, *DAB2*, *NOTCH2*, *MFHAS1*, *SPARC*) [238-243], which has recently been described to act as a barrier for the exit from pluripotency to endoderm formation (Fig. 31D) [217].

Phosphorylation levels of JUN-activating upstream kinase JNK are a strong indicator of JNK/JUN/AP1 signaling pathway activation [217, 244, 245], which we observed by increased relative amounts of pJNK in *LNC SOX17* depleted cells (Fig. 32A,B). Inhibition of JNK hyperactivity (JNK Inhibitor XVI) from day 3 of definitive endoderm differentiation only partially rescued the specification defect in *LNC SOX17* depleted cells (Fig. 32A,B) highlighted by CXCR4 expression recovery.

Deeper investigations of our transcriptional data revealed deregulated genes responsible for epithelial-to-mesenchymal-transition (EMT) (e.g., *CDH1*, *CDH2*, *SNAI1*, *VIM*, *CLDN1*) (Fig. 32E). Recent investigations have implicated a crucial role of EMT and its key driving TF SNAI1



**Fig. 33 *LNC SOX17* repressed endoderm lacks the ability to generate PP1 pancreatic progeny.** (A) Bright field images of PP differentiation cultures (upper panel) followed by IF staining for PDX1 (lower panel) of either sgCtrl or sgLNC SOX17 cells. Scale bars indicate 10  $\mu$ m. (B) IF staining quantification of overall (sgCtrl, n = 17657, sgLNC SOX17, n = 5279) PDX1+ population percentages (left) or PDX1 mean fluorescence intensity distribution in PDX1+ cells (right). Bar plot error bars indicate SD around the mean value and white dots represent mean values for the individual replicates (n = 10). Lines of the violin plot indicate interquartile range around the median value and white dots represent median values for the individual replicates (n = 10). (C) qRT-PCR showing the expression of *LNC SOX17* (left panel) and *PDX1* (right panel) during PP differentiation of sgCtrl and sgLNC SOX17 cells. Fold change is calculated relative to the 18s housekeeping gene. Symbols indicate mean values and error bars indicate SD across independent experiments (n = 3). (D) PCA of the 1000 most variable genes between sgCtrl and sgLNC SOX17 in PSCs and PP as measured by RNA sequencing. Gray dashed arrows indicate the two divergent transcriptomic trajectories. (E) Heatmap showing row-normalized z-scores of PP specific marker genes in sgCtrl and sgLNC SOX17 EN cells as measured by RNA sequencing at day 9 of differentiation. Columns were ordered by hierarchical clustering (represented as tree above the heatmap). Note the reduced expression of PP master transcription factor PDX1 in sgLNC SOX17 as compared to sgCtrl. (Landshammer A., Kretzmer H. and Bolondi A. contributed to this figure)

for the formation of hESC derived definitive endoderm[246]. Testing the hypothesis of potentially altered EMT due to loss of *LNC SOX17*, we performed immunofluorescent stainings for ECAD, NCAD, and VIM, which confirmed our transcriptional data and a retention of an epithelial signature in *LNC SOX17* depleted EN cells (Fig. 32F,G).

Finally, we evaluated if *LNC SOX17* repressed EN cells have lost their potential to further differentiate into pancreatic progenitor (PP) cells [247]. Immunofluorescent stainings identified a very distinct PDX1<sup>+</sup> population in cultures of control cells after nine days of directed differentiation. PDX1<sup>+</sup> cell fractions as well as expression levels were notably reduced in *LNC SOX17* depleted cells (Fig. 33A,B). In addition, transcriptomic analysis of differentiated control and *LNC SOX17* depleted PP cells indicated a substantial gene expression difference, including the specific downregulation of pancreatic progenitor marker genes [247] (Fig. 33C-E).

Taken together, our data highlight the importance of *LNC SOX17* for the formation of definitive endoderm its established and maintained transcriptome and its subsequent differentiation potential.

## 5 DISCUSSION

During development of a multicellular organism, a cornucopia of molecular processes are required to guarantee accurate gene regulation. Controlling gene expression tightly in space and time is key for the constitutive patterning, polarization and cell fate specification throughout the commitment of highly diverse lineages. At the nuclear level, this very complex regulation is in parts highly dependent on 3D chromatin organization, *cis*-regulatory elements (CREs) and long non-coding RNAs (lncRNAs). In the context of 3D chromatin organization, CCCTC-binding factor (CTCF) loop domains facilitate enhancer-promoter proximity and restrict inappropriate physical contacts. Nevertheless, the impact of such domains as architectural microenvironments on gene regulation remains controversial[89, 248]. The long non-coding transcriptome, however, is known to act in various and very diverse *cis*- and *trans*- acting ways via a plethora of mechanisms. Yet, the influence of individual lncRNAs and their genomic loci on gene control is still elusive, as indicated by the database LNCipedia (version 5.2), suggesting ~5% (49,372 high-confidence lncRNA genes, 2,482 manually curated lncRNA articles) of identified human lncRNAs to be functionally reported[249]. In this work I aimed to analyze this ambiguity in detail specifically at the domain of the *SOX17* locus. *SOX17* expression pattern in combination with the associated molecular phenotypes suits this locus perfectly for such an approach.

In the scope of our first published study (Wu H.J. and Landshammer et. al) and by dissecting the role of single-gene CTCF-loop domain boundaries, we were able to correlate changes in the spatial organization at the locus with *SOX17* gene expression analysis and the resulting differentiation phenotype. In contrast to other studies, this revealed the mere loss of insulation, especially of single genes isolated within CTCF-loop domains so called topologically isolated genes (TIGs), to be sufficient for developmental misexpression. These results further extended our understanding of genome architecture with respect to transcriptional output for the development of human definitive endoderm. In the scope of our second unpublished study (Landshammer A. and Bolondi A. et al), we dissected the *cis*-regulatory role of the *SOX17* locus associated DMR. We found *SOX17*-DMR to consist of a distal enhancer element, *eSOX17* and a promoter element *pLNC**SOX17* giving rise to a novel long non-coding RNA (lncRNA) transcript, *LNC**SOX17*. By studying both elements and the lncRNA, we could correlate their presence and activity with *SOX17* gene expression and associate the resulting phenotypes by various loss-of-function (LOF) approaches. Moreover, our studies highlight the importance of *LNC**SOX17* for the overall formation of definitive endoderm, its transcriptomic integrity and downstream differentiation potential. Our results not only extended the catalog of known enhancers and lncRNAs, but also expanded our knowledge of their relevance for early human development during gastrulation and the formation of definitive endoderm.

## 5.1 Topological isolation and its regulatory importance for developmental genes as *SOX17*

Genome organization and 3D chromosomal structural domains are crucial for precise gene control but the functional relevance of their boundaries on developmental processes is complex and remains insufficiently understood. Hence, resolving the relationship between 3D genome organization and its regulatory programs is of broad interest. Several studies have demonstrated the functional relevance on gene regulation of diverse systems by Mediator-Cohesin loops, mostly enabling enhancer-promoter interactions[[109](#), [113](#), [203](#), [250-254](#)]. CTCF Cohesin loops, which majorly work as insulators, have similarly been suggested to constrain enhancer-promoter interactions to guarantee proper gene expression patterns[[112](#), [137](#), [255-257](#)]. In our published study by Wu H.J. and Landshammer A. et al, 2021, we describe one aspect of genome organization that may facilitate the precise temporal and spatial control of key developmental regulators in human ESC differentiation. Our model is supported by functional data showing that disruption of a CTCF loop domain boundary of an “isolated” gene can strongly impact the lineage commitment of pluripotent cells exemplified at the *SOX17* locus.

We generated deeply sequenced Hi-C data for human pluripotent stem cells (hPSCs) which together with CTCF-ChIPs revealed CTCF-CTCF loop domains, highly overlapping with insulated neighborhoods, earlier identified by Cohesin ChIA-PET data in primed human embryonic stem cells (ESCs)[[203](#)]. Interestingly, although insulated neighbourhoods highly overlap with our identified CTCF loop domains they differ in domain number (~24,000 CTCF loops / ~13,000 insulated neighborhoods), median domain size (340 kb for CTCF loops / 190 kb for insulated neighborhoods) and median number of genes per domain (1 gene for CTCF loops / 3 genes for insulated neighbourhoods). These numbers may vary depending on different identification techniques, target proteins, computational calling methods and assumptions made for filtering genomic data[[258](#)]. To our surprise, among CTCF loop domains we found a big fraction to contain only a single gene in a non-random fashion, supporting the notion of topological “isolation” of a certain subset of genes in human ESC genomes. Therefore, we term these single-gene loop domains, and the genes contained as such topologically isolated genes (TIGs), highlighting their domain-specific importance in ESC genomes.

Intrigued by the number of isolated genes we had a closer look into stability of CTCF loop domains beyond ESCs and find that their boundaries are largely preserved throughout the process of ESC differentiation. In particular, we found CTCF single-gene loop domain boundaries throughout three germ-layer differentiation to be even more preserved than others. A number of studies tested whether 3D chromatin architecture is important for gene regulation

by depleting CTCF or Cohesin protein[120-124]. Surprisingly, neither the depletion of CTCF nor of Cohesin had a strong effect on gene expression[121-123]. The observed weak effects on gene regulation challenged the concept that 3D genome organization is essential for enhancer-promoter interaction and gene regulation. It is important to keep in mind, however, that these experiments measured the effects on gene regulation *in vitro* in cell culture in a steady state and thus did not test the relevance of 3D chromatin architecture for gene regulation during more tightly restricted regulatory events, such as development, differentiation, or cell signaling. Thus, while ruling out an essential role in enhancer-promoter interaction, these studies point to a more complex, multilayered effect of 3D chromatin structure on gene regulation. In concordance with our findings and the idea of a permissive model of enhancer-promoter proximity[125], this would underline the need for temporal precision of gene regulation to ensure fast activation and safeguard proper developmentally required levels of gene expression, sustained by CTCF loop domains.

Although many features of 3D genome organisation are established at or during zygotic genome activation (ZGA), zygotic transcription itself is generally not required, at least for the establishment of topologically associating domains (TADs) and their boundaries[259-264]. The exception is human embryos. One reason for this might be that CTCF is not maternally provided in human embryos and must be expressed from the zygotic genome. Different than in mouse sperm cells[265], human sperm cells do not express CTCF either, also lacking the establishment of TADs[204]. TADs are gradually established during embryonic development following human fertilization[204]. As TADs, A/B compartmentalization is lost in human embryos at the 2-cell stage and is re-established during embryogenesis. Blocking ZGA inhibits TAD establishment in human embryos but not in mouse[204]. During mouse development, depletion of zygotic and maternal CTCF results in embryo lethality[266, 267]. While TADs are formed, they have reduced insulation. In human, CTCF is not maternally inherited, and zygotic CTCF is required, but not sufficient, for TAD formation[204]. However, providing CTCF in embryos where transcription has been blocked is not sufficient to rescue genome organisation [204], suggesting that transcription has an additional role. Our results indicate that CTCF has a key role in the establishment of 3D chromatin structure during human embryogenesis. Nevertheless, we found CTCF loop domains to be more resistant in early human embryonic ZGA inhibition experiments[91, 204], supporting the idea of once established within ZGA, single-gene domain boundaries are maintained independently of ZGA and of CTCF expression and might thus be more stable and robust across diverse cellular processes. It might be that other factors play a role at these domain boundaries as e.g. Cohesin release factor WAPL[120], structural regulator YY1[268] or Cohesin[122] itself, potentially being involved in preservation of single-gene loop domains and their implied functional relevance across ESC differentiation.

Curious about the stability of established single-gene loop domains, we sought to identify genetic features and found developmental regulators to be enriched in single-gene loop domains over multi-gene and no gene loop domains. These results motivated us to test evolutionary conservation of these boundaries as a proxy for preserved function. We find that CTCF boundaries of single-gene loop domains appear highly conserved across different mammals, in particular across placental mammals and marsupials. These findings suggest that motifs of single-gene loop domains are functionally important elements that undergo natural selection, more than compared to multi-gene or no gene loop domains. From the gene-perspective we find that CTCF sites of developmental regulator domains are more conserved than those of other genes, which implies additional functional importance. Intrigued by the gene ontology within single-gene loop domains, we refined the list of developmental regulators (DRs) to early developmental regulators (eDRs). eDRs are differentially expressed among the 4 states of pluripotency and the three germ-layer derivatives as e.g., *SOX17* exclusively during the formation of definitive endoderm. Interestingly, all CTCF loop domains show much higher enrichment in eDRs than DRs over all other genes (AG), especially when only containing one gene per domain. These single-gene loop domains surprisingly also contain more CREs per gene, such as enhancers and lncRNAs, than multi-gene or no gene domains, which again hints towards a requirement for tight and dynamic regulation of those genes within CTCF loop domains.

How important TIGs and their topological isolation may be in a disease context has recently been shown by several studies[269, 270]. As algorithmically defined by Long H.S. et al, TADs for instance can be classified into two functionally different groups, those which are bound by CTCF and those which are not. The authors find no association between genes sharing the same CTCF TADs in regards of increased co-expression or functional similarity, other than that explained by linear genome proximity[266]. However, their data hints towards a similar direction as of TIGs, suggesting that genes in TADs on their own are less tolerant to mutations, hence the tight control of these genes to be highly relevant in a mutation-associated disease context. Interestingly, this observation is highly in concordance with our data identifying depletion of common variants (allele frequency >1% in the population) especially in CTCF loop domain boundaries of constitutive loop domains across different tissues, indicating that boundaries of constitutive CTCF loop domains, if altered, are subject to purifying selection (Wu H.J. and Landshammer et al., 2021 Fig. 6e)[91, 269]. The idea to isolate developmental genes by CTCF loop domains and the functional relevance of their boundaries in regards of purifying selection, underlines the regulatory importance of TIG boundaries to safeguard gene regulation. That said, another TIG, *GPR101* and its disease related altered 3D chromatin architecture has recently been associated with X-linked acrogeria (X-LAG)[271]. The authors show that Xq26.3 duplications, finally missing natural boundary constrain, leads to

massive pituitary tumoral expression of the novel growth hormone (GH) regulator GPR101, acquired by ectopic enhancer regulation[271]. Similar observations have been made when reintroducing genetic alterations of perturbing CTCF boundaries, similar to their somatic mutations in patients, which led to gene de-regulation of insulated neighborhoods containing T-ALL oncogenes. Interestingly, among them are TIGs as e.g., *LMO1* and *NOTCH1*, both known to be highly associated with cancer pathogenesis[256]. Our obtained results suggest that CTCF-CTCF single-gene loop domains with their highly conserved and stable boundaries are of regulatory importance for proper gene-control during development. Topological “isolation” of developmental regulators or TIGs e.g. *SOX17*, further complements the field of 3D chromatin architecture perspective wise and supports functional importance for proper gene-control in development and disease.

TIG *SOX17* is a well known key developmental regulator involved in early developmental processes[32, 34]. It is suggested as a tumor suppressor in several cancer types[30, 194, 195] but besides its functional role at the protein level, little is known about its own regulation, especially in definitive endoderm. So far, only its CpG methylation landscape and the association with tissue-specific TFs in definitive endoderm[21, 29] has been described but its 3D chromatin architecture and regulation in *cis* during the formation of definitive endoderm at the *SOX17* locus is barely understood[180-182]. Intrigued by the question if TIG *SOX17*'s strong CTCF-boundaries are functionally relevant for gene-control, we performed a functional study on the *SOX17* locus. This demonstrated that disruption of a CTCF loop domain boundary leads to *SOX17* deregulation and a comprised definitive endoderm phenotype impacting cell lineage commitment of human induced pluripotent stem cells (iPSCs). In particular, boundary perturbation at the *SOX17* locus causes decreased interactions between DREs and the *SOX17* promoter. Moreover we could exclude the possibility of ectopic enhancer adoption/hijacking of *SOX17*-DREs by genes within the upstream loop domain. Interestingly, transcriptomic profiling revealed differentiating cells to be stalled in an earlier “mesendodermal like” state of differentiation, reversible by ectopic *SOX17* expression. Enhancer adoption/hijacking after boundary perturbation has been widely observed and extensively studied in both development and tumorigenesis[137, 206, 256, 257, 271]. Nevertheless, it remains elusive if in case of the *SOX17* locus, certain enhancer-promoter compatibilities are present and required for gene-control. One current model suggest, promoters might have sequence-encoded preferences for certain enhancers, for example mediated by interacting sets of TFs or cofactors[272]. This “biochemical compatibility” model has been supported by observations at individual human promoters and by genome-wide measurements in *Drosophila melanogaster*[273-279]. Interestingly, recent genome wide combinatorial compatibility investigations revealed, most enhancers to activate any promoter by similar amounts. Not surprising is the fact that intrinsic enhancer and promoter activities combine multiplicatively to



determine RNA output. However, two classes of enhancers and promoters show preferential effects. Housekeeping gene promoters for instance contain built-in activating motifs for factors such as GABPA and YY1, which in turn decrease the responsiveness to distal enhancers. Nevertheless, promoters of differentially expressed genes, do lack these motifs and show stronger responsiveness to enhancers, suggesting a multiplicative model tuned by enhancer-promoter classes to control transcription[280]. Interestingly, the fused *SOX17* upstream CTCF loop domain contains housekeeping genes, expressed (*ATP6V1H*, *TCEA1*, *LYPLA1*, *MRPL15*) or not expressed (*RGS20*) across differentiation. Hence, one may speculate their promoters to potentially have built-in motifs for generic factors decreasing the responsiveness for the boundary-perturbation accessible distal *SOX17* enhancer and potential “shadow enhancer”. At the same time these built-in motifs may not be contained within the *SOX17* promoter, which is in fact a differentially expressed gene, just induced upon the formation of definitive endoderm. Therefore, our findings support the above-mentioned model but further investigations, e.g. motif-enrichment analysis of the respective promoters may be required.

With our model we demonstrated that boundary perturbation at the *SOX17* locus did not induce enhancer adoption by other genes but cause loss of enhancer regulation of its endogenous target. Taken together, our results suggest a dual function of topological insulation – the boundary interaction not only constrains enhancer activity within the domain, but also facilitates enhancer-promoter interaction by bringing them into physical proximity. This observation also implies the existence of diverse mechanisms of topological insulation[281], which need to be dissected further.

## 5.2 The *SOX17* DMR is comprised of two distinct CREs of diverse function

Concomitant with CTCF loop domain boundaries, promoters and enhancers are key elements regulating gene-control and precise spatiotemporal expression of genes. As previously shown for the *SOX17* locus, its differentially methylated region (DMR), is functionally implicated in contacting and enhancing *SOX17* expression during the formation of definitive endoderm[29, 91]. Therefore, it was of interest to further dissect the DMR, to define which regions in particular would facilitate contact and how its epigenetic, transcriptional and TF-landscape would look like. In our second, unpublished study by Landshammer A. and Bolondi A. et al, 2022 *in review*, we studied the *SOX17* DMR in more precise detail. In the scope of this study, we attempted to answer exactly those questions and explored the characteristics and function of the *SOX17*-DMR during the formation of definitive endoderm. As shown previously in human ESCs, the *SOX17*-promoter and its DMR differ quite in the acquisition of the repressive histone mark H3K27me3 and the active transcriptional histone mark H3K4me3 upon endoderm differentiation. The *SOX17* promoter has been shown to be bivalently[282] marked by low H3K4me3 and high H3K27me, but when differentiating into definitive endoderm this epigenetic

state changes into an active state, depicted by high H3K4me3 and low H3K27me levels respectively[29]. Bivalent domains have been reported to coincide with developmental TF genes and are proposed to silence developmental genes in embryonic stem cells (ESCs) while keeping them poised for activation[282]. This mechanism would allow undifferentiated cells to quickly respond to respective signaling in order to activate the required transcriptome for development into the required cell fate. Interestingly, for the *SOX17* DMR, the bivalent state is exclusively acquired upon the formation of definitive endoderm by both marks decorating the locus, while being absent in ESCs[29]. High levels of H3K27me3, as suggested by the literature, are unusual for active enhancers which are generally reported to be marked by H3K4me1 and H3K27ac[283, 284]. This indicated that the *SOX17*-DMR may be both, an active enhancer, and a bivalent promoter at the same time, potentially primed for downstream endoderm arising tissues to acquire different states of activity.

Intrigued by the actual identity of the *SOX17*-DMR, we in depth epigenetically profiled the region and dissected its function. We found the *SOX17*-DMR to be comprised of two distinct sites, one being the cognate fully functional and tissue specific *SOX17* distal enhancer, namely *eSOX17* with its core *eSOX17.2*, exclusively physically interacting with *pSOX17* in a tissue specific manner. Genetic ablation studies of *eSOX17.2*, indicated strongly decreased endoderm differentiation induction highlighted by reduced *SOX17*<sup>+</sup> cell fractions. After induction following day 3 of differentiation, decreased *SOX17*<sup>+</sup> cell fractions were able to finally keep up, indicated by recovering percentages as compared to wild type. This suggests that *eSOX17.2* is functionally relevant for the tissue-specific induction of *SOX17*. However, the terminal differentiation-outcome was found unaltered (*CXCR4* fractions at day 5) compared to the wild-type scenario. These results may be explainable by either the functional and redundant role of *eSOX17.1* or potential “shadow enhancers” present within the loop-domain, compensating a lack of *eSOX17.2* but temporally being less efficient. “Shadow enhancers” are CREs of seemingly redundant regulation and function which drive gene expression in overlapping expression patterns. Recent studies have pinpointed them to be remarkably abundant, controlling most developmental gene expression from invertebrates to mammals. Hence, they might provide crucial mechanisms for gene expression buffering, giving robustness against mutations of regulatory regions for genes implicated in human disease. In addition, evolutionary conservation and prevalence of “shadow enhancers” underscore their key role in emerging metazoan gene regulatory networks[285]. In regards of the strong observed CTCF loop domain boundary perturbation phenotype, this may be an explainable scenario since disruption of the loop domain can be seen equal to the simultaneous deregulation of all DREs within the *SOX17* loop domain. Nevertheless, these assumptions still need to be experimentally tested and further validated by distinct and combinatorial enhancer manipulations within the *SOX17* loop-domain.

Curious about the second region within the *SOX17* DMR more upstream of *eSOX17*, we attempted further characterization and epigenetic dissection. Most surprisingly, Chrom-HMM indicated a poised promoter signature for the more upstream region within the DMR, which we named *pLNC**SOX17*. Active promoters of PolIII driven gene-bodies bear an actual sequence-orientation, are marked by H3K36me3/H3K4me3, can transcribe genes at low levels independently of enhancers and produce 5' capped and 3' polyadenylated RNAs[286]. In comparison, active enhancers which have no orientation, marked by H3K4me1/H3K27ac, can only initiate PolIII induced transcription in collaboration with promoters and create often short, undirected, and post-transcriptionally unmodified eRNAs[283, 284]. When validating *pLNC**SOX17*s promoter identity by promoter luciferase activity assays, we identified *pLNC**SOX17* to bear definitive endoderm tissue-specific activity. Luciferase activity assays of constructs with inverted orientation in the designated vector system or alternatively, original constructs in the respective undesignated vector system shall ultimately confirm this identity of *pLNC**SOX17* and *eSOX17* in future experiments. LncRNA genes are very often enriched at enhancer elements and their expression is highly tissue-specific[160, 287-289], as it has been shown for many enhancer elements too[290, 291]. One approach to identify long intergenic non-coding RNAs (lincRNAs) is testing for distinct chromatin-states in combination with poly(A) RNA sequencing to discover discrete transcriptional units intervening known protein-coding loci [292]. Doing so, we were surprised to discover a so far unknown and non-annotated 22 kb long RNA PolIII transcript, which we named *LNC**SOX17*. Due to the proximal localization and tissue specificity of lincRNA genes and enhancer elements, it is important to distinguish between lincRNAs transcribed from enhancers (sometimes referred to as enhancer lincRNAs or e-lincRNAs) and another species of non-coding RNAs produced at enhancers, termed enhancer RNAs (eRNAs)[293]. Although the two terms are often conflated, and although some enhancers produce both lincRNAs and eRNAs, the main distinctions between eRNAs and e-lincRNAs are size, stability, biogenesis/processing and their transcriptional directionality.[293] In light of our collected evidences (size, histone-profile, polyadenylation, directionality and identification/characterization of *pLNC**SOX17*) we were convinced *LNC**SOX17* in fact to be a novel lincRNA, driven by its promoter *pLNC**SOX17*. Final epigenetic and TF binding profiling along the entire *LNC**SOX17* locus confirmed this evidence and led to the conclusion that indeed, *eSOX17* is the only regulatory element with enhancer identity overlapping with the *LNC**SOX17* locus. Hence, in contrast to *pLNC**SOX17*, *eSOX17* with its core *eSOX17.2* is the only element being a *bona fide* distal *SOX17*-enhancer during the formation of definitive endoderm. Moreover, we suggest *LNC**SOX17* not to be a e-lincRNA, since being driven by its very own cognate promoter *pLNC**SOX17*.

### 5.3 *LNC*SOX17 is dependent on SOX17 and does not regulate its locus in *cis*

Characterizing *LNC*SOX17 further, we find this novel lincRNA to be present in similar developmental stages of various vertebrates. In human iPSC derived EN, we identify *LNC*SOX17 to be a *bona fide* lincRNA, highly tissue-specific and uncoupled from *SOX*17 expression across downstream endodermal tissues. Transcriptomic single-cell RNA (scRNA) sequencing data of the human gastrulating embryo[17] revealed *LNC*SOX17 to be exclusively expressed in endoderm and processed highly similar to *in vitro* derived endoderm *LNC*SOX17 cDNA, confirmed by MinION long-read sequencing and validated by 5'/3' rapid amplification of cDNA ends PCR (RACE-PCR). We find *LNC*SOX17 to be localized mainly within the nuclear space, its open-reading-frames (ORFs) to bear no coding potential and identify 2 main isoforms besides most RNA associated to the locus being “sloppily” spliced/transcribed. Taken together our findings report *LNC*SOX17 to be a *bona fide* novel lincRNA, having all characteristics as commonly described lincRNAs[294], present in definitive endoderm of the early gastrulating human embryo. Therefore, our study further complements the catalogue of unknown functionally relevant endodermal lincRNAs and extends our understanding of lincRNA biology during early human gastrulation. Long non-coding RNAs have widely been implicated in development over the last decades[295] and were shown to be crucial for development of the human embryo in various instances[296-298]. Nevertheless, their modes of action and how they govern gene-control are quite diverse and highly dependent on their protein-interaction partners[299, 300]. So far, very little is known about the broader scale of lincRNAs and their biological relevance, especially in early human gastrulation. One of the germ-layer – definitive endoderm – is a highly relevant tissue within the embryo since it gives rise to several crucial digestive and detoxifying organs as the pancreas, liver, and the gastrointestinal tract. Hence, there is great interest in understanding the pathways that regulate the induction and specification of this germ-layer. Recent CRISPR interference/activation (CRISPRi/CRISPRa) screens identified dozen lincRNAs in definitive endoderm[301], but only 5 individual case-studies have been describing endodermal lincRNAs in more detail and giving a glimpse into their mode of gene-regulation. These lincRNAs, namely *DEANR1*, *DIGIT*, *GATA6-AS* and *LINC00458* have been found crucial for the formation and the regulation of proper definitive endoderm, by controlling this germ-layer derivative in either a *cis*[302, 303] or *trans*[66, 231, 304] acting manner.

Hence, we asked whether *LNC*SOX17 may be involved in *SOX*17 *cis*-acting gene-control. To do so, we utilized a constitutive CRISPRi system to epigenetically silence the *LNC*SOX17 locus, targeting its promoter *pLNC*SOX17. Exploring the function of genes, one way is to disrupt their expression through repression. The dominant tool for programmed knockdown of

mRNAs is RNA interference (RNAi)[305]. However, RNAi comes along with several problems e.g., off-target effects, which can be especially confounding in the context of cellular differentiation and cell identity[306-308]. Additionally, because RNAi is mediated by cytoplasmic argonaute proteins, gene silencing through this approach is best suited to depletion of cytosolic mRNA targets. Hence, the chosen CRISPRi approach seemed beneficial over RNAi approaches. Utilizing CRISPRi mediated repression of the *LNC*SOX17 locus, we were also able to simultaneously prevent active ongoing transcription.

Even though *LNC*SOX17 was strongly repressed, we did not observe changes of the 3D chromatin architecture at the locus, neither did we identify altered SOX17 occupancy locally and throughout the genome. Most interestingly, we did not observe changes in SOX17 RNA and protein expression due to loss of *LNC*SOX17. Different then to *LNC*SOX17, *DEANR1* and *GATA6-AS* are linearly proximal to their early developmental regulator TFs[302, 303]. Although *DEANR1* is located in proximity to *FOXA2* it is driven by its very own promoter, whereas *GATA6-AS* and *GATA6* are both driven by one bidirectional promoter[302, 303]. *DEANR1* and *GATA6-AS* regulate both their proximal TF genes in *cis* by SMAD2/3-tethering to their target promoter elements[302, 303]. In the case of *GATA6-AS*, active transcription of the lncRNA locus increases the overall activity of the entire region, while the lncRNA itself in a complex with SMAD2/3 governs *GATA6* expression in a *cis*-activating manner[281]. *DEANR1* lncRNA also associates with SMAD2/3 and the RNA-protein complex further facilitates physical looping between the *DEANR1* locus and the *FOXA2* promoter tested by RNA-DNA fluorescent *in situ* hybridization (FISH) [280]. These results may be misleading, regarding the close proximity of nascent RNA-production and the *FOXA2* promoter. Hence, further experimental validation may be required. Besides the physical interaction of *e*SOX17-*p*SOX17 leading to proximity of the *LNC*SOX17 and *SOX17* locus in 3D, different to *DEANR1* and its regulation of *FOXA2*, ablation of *LNC*SOX17 transcription and absence of its lncRNA has no impact on *SOX17* gene-regulation and its expression levels. Therefore, we hypothesize an absent *cis*-acting mechanism of *SOX17* by *LNC*SOX17 and a different type of regulation for *LNC*SOX17 compared to mechanistically studied endodermal *cis*-regulating lncRNAs. Other than the interaction of *DEANR1* and *GATA6-AS* with SMAD2/3, our results suggest that *LNC*SOX17 may not be associated with SOX17 protein and involved in facilitating binding DNA by SOX17 locally and globally. Nevertheless, identifying specific protein interaction partners may be crucial to fully understand *LNC*SOX17 mode of action and will give further insights into its regulatory mechanism.

Since *LNC*SOX17 does not to regulate its upstream gene *SOX17* in *cis*, we hypothesize a potential *trans*-acting mechanism by *LNC*SOX17 in regards of the resulting phenotypes (discussed below). lncRNAs, primarily in the nucleus near their site of transcription are found to exert transcriptional regulation of a proximal gene (*Cis* regulation of proximal loci in 3D) as

for instance *DEANR1*[302] in definitive endoderm or others e.g., *Xist*[309] or *HOTTIP*[310],[311]. Alternatively, localization of lncRNAs across the nucleus instead indicates rather transcriptional regulation of distal genes (*Trans* regulation of distal loci in 3D) as for instance found for *DIGIT*[231] in definitive endoderm or other lncRNAs e.g., *Firre*[312] or *NEAT1*[313]. Interestingly, RNA fluorescence *in situ* hybridization (RNA-FISH) experiments revealed distribution of exonic *LNC SOX17* across the entire nuclear space, supporting our hypothesis and suggesting that *LNC SOX17* may regulate the genome potentially in *trans*, during the formation of definitive endoderm. Additionally, temporal CXCR4 profiling of *LNC SOX17* repressed and control cells reveal induction of CXCR4 at day 3 (induction time-point for *LNC SOX17*) and a lack of CXCR4 maintenance for day 4-5 upon loss of *LNC SOX17*. These results give first indications towards definitive endoderm gene deregulation upon loss of *LNC SOX17* concomitant with unaltered *SOX17* levels, potentially regulated in *trans* directly by *LNC SOX17*. Nevertheless, this outcome could also be the consequence of failed definitive endoderm differentiation, and in fact an egg-hen problem. If *LNC SOX17* does regulate endodermal genes (e.g. *CXCR4*) in *trans*, one could speculate a maintenance mechanism to guarantee constant expression of *trans*-genes during developmental progression. How relevant such a maintenance may be in terms of development, could potentially be tested by perturbing *LncSox17* expression in other model systems, e.g. *Mus musculus*. These results may give new insights to associate molecular/cellular phenotypes and link these to developmental processes.

The lncRNA *DIGIT* and its locus is so far the most comprehensively described lncRNA in definitive endoderm[66, 231]. *DIGIT* and its associated protein-coding gene *GSC*, are activated in *cis* by a 5kb proximal enhancer element downstream of *DIGIT*[66]. Deletions of the SMAD3 occupied enhancer leads to miss-expression of *DIGIT* and *GSC*, identifying it as the CRE regulating both genes at the locus specifically in endoderm[66]. Small hairpin RNA (shRNA) and anti-sense locked nucleic acid (LNA) knockdown of *DIGIT* led to compromised definitive endoderm too, highlighting the actual *DIGIT* transcript as CRE-downstream cause of the differentiation phenotype[66]. Similar results were obtained by early polyadenylation-reporter knock-in perturbations, 44 bp downstream of the *DIGIT* TSS, showing the importance for the lncRNA itself and not active transcription of the *DIGIT* locus causing endoderm differentiation failure. This result was shown to be consistent in mESC derived endoderm targeting the ortholog locus of *Digit*[66]. To identify if RNA or active transcription is involved in regulation of a gene, different models have been established[314]. To reduce RNA levels without impacting transcription of the locus, early transcriptional termination has been proposed, which can be achieved by CRISPR/Cas9 based knock-in of donor DNA sequences bearing early polyadenylation signals after the transcriptional start site (TSS) of a lncRNA[230, 314]. To rule out any *cis*-regulatory effect at the *SOX17* locus and to confirm our earlier CRISPRi based

results, showing *LNC*SOX17 RNA to cause the observed CXCR4 deregulation phenotype, we performed CRISPR/Cas9 based knock-in of an alternatively transcribed sequence followed by strong transcriptional termination signals. qRT-pPCR of early terminated *LNC*SOX17 reporter iPSCs highlighted a definitive endoderm specific absence of *LNC*SOX17 leading to an exact similar, temporal deregulation of CXCR4 maintenance from day 3 of differentiation onwards, accompanied by unaltered SOX17 levels. These results pinpoint a phenocopy of the earlier CRISPRi model, strongly suggesting *LNC*SOX17 RNA itself causing the CXCR4 phenotype, similar to lncRNA *DIGIT*. Interestingly in comparison to *LNC*SOX17, all perturbation models for *DIGIT* show reduced levels of GSC<sup>[66]</sup>, a main driving TF of definitive endoderm formation in early gastrulation, which is not observed for *LNC*SOX17 perturbations in regards of unaltered SOX17 mRNA and protein levels. Ectopic GSC-rescue experiments in *DIGIT* early polyadenylation-reporter knock-in perturbations, nevertheless, could compensate the *DIGIT* ablation associated endoderm formation failure. Most surprisingly, this was consistently repeated by ectopic *DIGIT* rescue experiments which proved *trans*-activity of the proximal lncRNA and underline *DIGIT* transcripts to also regulate GSC gene control downstream of the CRE. To ultimately prove *trans*-regulation for *LNC*SOX17, ectopic lncRNA (over-)expression shall be able to rescue the CXCR4 phenotype, yet still part of future investigations.

Intrigued by SOX17 occupancy at *pLNC*SOX17 and the locus regulation in *cis*, we challenged a *vice versa* dependence of SOX17 for *LNC*SOX17 activation. *SOX17* gene-body ablation led to poor differentiation outcome accompanied by a lack of *LNC*SOX17 expression compared to wild-type or heterozygous deletions. This suggests that a lack of SOX17 and poor differentiation outcome leads to the complete absence of *LNC*SOX17 expression, which is concordant with SOX17 occupancy at *pLNSOX17*, proposing a SOX17-dependence for *LNC*SOX17 activation. Nevertheless, we cannot exclude the fact that SOX17 perturbed cells may not even differentiate far enough, potentially running again into an egg-hen problem of overall differentiation failure that does not acquire a state for potential *LNC*SOX17 activation. In that circumstance, a general definitive endoderm failure, which could further lead to the absence of endodermal gene expression upstream of *LNC*SOX17, may potentially cause lack of other TFs besides SOX17, to not induce *LNC*SOX17 expression. To ultimately confirm SOX17-dependence for *LNC*SOX17, one could diminish SOX17 protein levels inducible by protein degradation at differentiation day 3 onwards. If the established *LNC*SOX17 levels upon SOX17 degradation would further be maintained, SOX17 dependency for *LNC*SOX17 expression maintenance could be excluded. However, SOX17 dependency for *LNC*SOX17 expression maintenance would be strongly evident, if *LNC*SOX17 levels instead decrease upon SOX17 degradation. Further the question remains, what else than SOX17 exactly drives the expression of *LNC*SOX17 during the formation of definitive endoderm and guarantees its tissue-specificity. Previous investigations of our lab and others studying the interplay between

TFs across pluripotency and the three germ-layer, revealed TF co-binding relationships that allow conclusions on the cooperativity for activation of tissue specific CRE activity[[29](#), [179](#)]. Hence, *LNC SOX17s* tissue-specificity might be explained by cooperatively targeting of its promoter element *pLNC SOX17*, occupied by a specific set of TFs as shown earlier (SOX17, FOXA2, GATA6, GATA4). One could speculate that besides extremely high specificity, TF-cooperativity may prevent misexpression and developmental failure.

Taken together our results suggest absent *cis*-regulation of *SOX17* by *LNC SOX17* RNA/transcription. Yet, unclear if *LNC SOX17* RNA may be involved in the regulation of other regions throughout the genome in *trans*, it is certainly an interesting molecule and worth identifying its potential interaction partners and mode of action.

#### **5.4 Loss of *LNC SOX17* leads to an aberrant definitive endoderm phenotype**

Endodermal lncRNAs have been implicated in crucial control of their respective associated key TF-genes and the formation of that germ-layer[[66](#), [231](#), [302-304](#)]. Since loss of *LNC SOX17* was found to lack CXCR4 maintenance, it was of interest to know if and how development is generally altered, in particular the formation of definitive endoderm. Directed and random differentiation profiling of *LNC SOX17* repressed cells showed unaltered ectoderm (EC) and mesoderm (ME) formation, but a compromised ability of definitive endoderm (EN) formation. These results imply in fact an ongoing differentiation failure specifically in definitive endoderm but not the other germ-layer upon loss of *LNC SOX17*.

To improve our understanding of genome wide expression alterations due to loss *LNC SOX17* and identify affected genes of the endodermal transcriptome, especially from day 3-5 of endoderm differentiation (induction and initial presence of *LNC SOX17*), we carried out temporal transcriptomic profiling for day 0, 3, and 5 definitive endoderm in *LNC SOX17* repressed and control cells. Interestingly, we find little to no changes in transcriptomes of CRISPRi and control for day 0, and 3 differentiated cells, revealed by PCA-analysis of the most variable genes (data not shown). Nevertheless, transcriptomic differences start to arise from day 3 onwards at day 5. This suggests genome wide transcriptional changes to appear following loss of *LNC SOX17* by direct or indirect dependence. Since day 5 of differentiation indicates the strongest transcriptomic changes, we further investigated differentially expressed genes (DEGs). We find genes associated with definitive endoderm and Wnt-signaling downregulated especially from day 3 onwards. Day-by-day resolved qRT-PCR of selected key definitive endoderm marker genes showed a lack of activation from day 3 onwards in *LNC SOX17* repressed and depleted cells, which confirmed a transcriptomic dependence of *LNC SOX17* at differentiation day 3 onwards. However, global transcriptomics revealed pluripotency factors included (*NANOG*, *POU5F1*) to be upregulated due to loss of *LNC SOX17*,



supporting the notion of differentiation failure due to its loss. Altogether, our results indicate a causal relation between terminal endoderm differentiation failure due to loss of *LNC SOX17* as a global transcriptomic dependency, potentially rescuable by ectopic expression of *LNC SOX17*. To rescue loss of *LNC SOX17* in the most “natural” way, it is suggested to integrate the *LNC SOX17* locus (including *pLNC SOX17*) homozygously into a safe harbour (*AAV1* locus) of the human genome. Utilizing that strategy, one could prevent eventual silencing of the transgene at random integration-sites in the genome and guarantee its temporal definitive endoderm activation from day 3 on. Once integrated, it is also suggested to test transgene expressed RNA biogenesis, distribution, and tissue-specific expression of the transcripts, in order to mimick *LNC SOX17* expression as natural as possible. Yet still part of future investigations, these experiments should finally prove *trans*-function of *LNC SOX17*.

Recent investigations have been associating the JNK/JUN/AP1 signaling pathway and its hyperactivity with inhibition of definitive endoderm formation[217]. The authors postulate a JUN-dependent inhibition of SMAD2/3 re-configuration for binding to pluripotency TFs versus definitive endodermal TFs. To test if loss of *LNC SOX17* may be associated with JNK/JUN/AP1 hyperactivity we investigated our DEGs at day 5 for CRISPRi and control. Indeed, we find JUN/AP1 target genes to be significantly upregulated due to loss of *LNC SOX17*. Validation of these results revealed JNK hyperactivity exclusively at day 5. Interestingly, chemical inhibition of JNK partially rescued our previously observed CXCR4 phenotype, which suggests that a proportion of cells may still respond to SMAD2/3 re-configuration via the JNK/JUN/AP1 pathway even though not expressing *LNC SOX17*. This, could potentially be an indirect effect of the JNK/JUN/AP1 pathway, uncoupled from definitive endoderm regulation by *LNC SOX17* and needs to be investigated further. Interestingly, an earlier study postulated similar importance of the epithelial-to-mesenchymal transition (EMT) process, by which chemical inhibition of TGF- $\beta$  or genetic ablation of EMT inducing-TF *SNAI1*, further leads to inhibition of definitive endoderm formation[246]. EMT which is a crucial process allows epiblast cells during gastrulation within the primitive streak to migrate and displace hypoblast cells to form the final definitive endoderm germ-layer[5, 6]. qRT-PCR and IF-stainings of *LNC SOX17* repressed and control cells indeed indicated EMT marker deregulation and a retained epithelial over mesenchymal phenotype, due to loss of *LNC SOX17*. As *CXCR4*, *SNAI1* could potentially be one of multiple *LNC SOX17* *trans*-regulated target genes. JUN/AP1 target genes, e.g. *SNAI1*[197] may potentially overlap partially with *LNC SOX17* target genes, which would explain why hyperactivation of the JNK/JUN/AP1 pathway may in fact partially rescue the CXCR4 phenotype. Under these circumstances, once activated by the JNK/JUN/AP1 pathway, one possibility might be that *LNC SOX17* controls target gene maintainance as found for *CXCR4*. Altogether, the underlying results strongly highlight the importance of *LNC SOX17* and associate its loss to EMT-failure and JNK hyperactivity, both leading to inhibition of

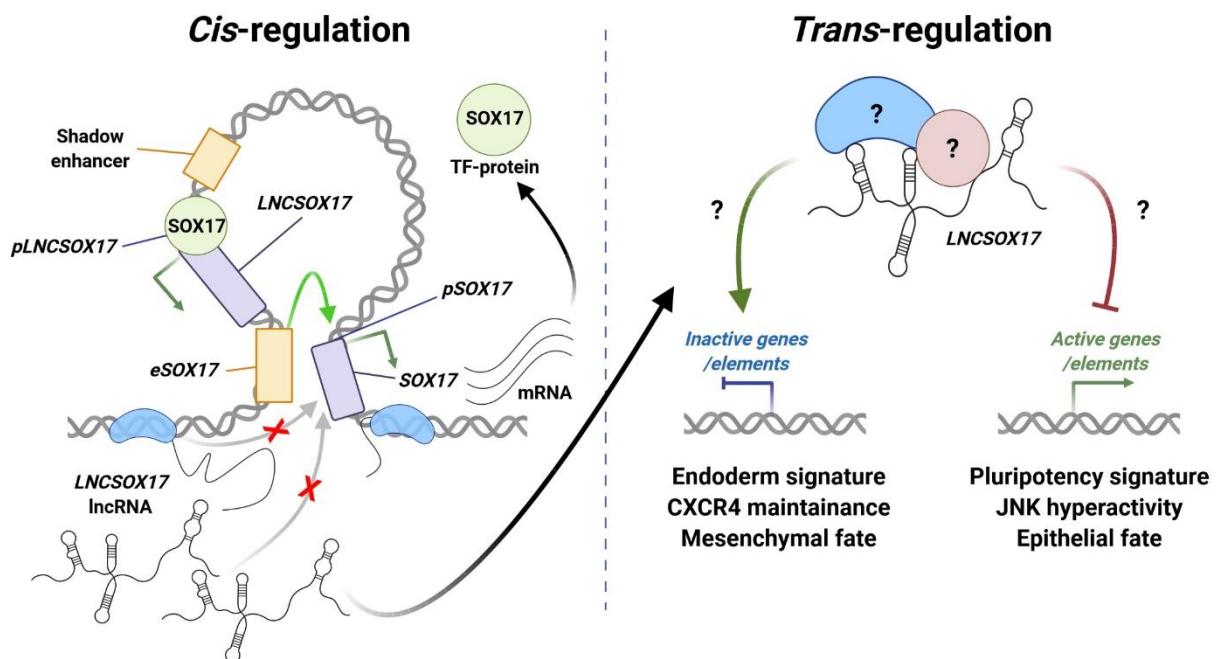
definitive endoderm formation. Additionally to the outlined data, we found *LNC SOX17* ablated definitive endodermal cells to also express high levels of visceral endodermal marker genes (*DAB2*, *PVR*, *HAPLN3*, *GSTM3*, *SLC6A8*, *STAT3*, *SEMA4A*, *SLC16A10*, data not shown)[317, 318]. Although definitive endoderm per definition is purely epiblast derived[317], one may speculate that *LNC SOX17* might be involved in safeguarding the definitive endoderm fate to constrain certain cell-plasticity. As for instance, *Sox17* mutant mouse embryos show besides foregut deformations, deficiency in midgut and hindgut development, concomitant with definitive endoderm cells exhibiting a visceral endoderm-like morphology[32]. Further investigations highlighted *Sox17*'s importance to maintain a definitive endoderm program, allowing cell egression during mouse gastrulation, indicated by definitive endodermal progenitors retained within the mesoderm layer[319]. Since mesenchymal-to-epithelial transition (MET) was postulated to be key allowing cell egression[319], one could imagine a similar scenario for *LNC SOX17* ablated cells, which fail to perform proper epithelial-to-mesenchymal transition (EMT), potentially being detrimental for the formation of definitive endoderm during human gastrulation. This hypothesis could be tested in mouse embryonic complementation assays to find out if and to which type of tissue *LNC SOX17* ablated cells would participate in, being part of future investigations.

*SOX17*-null mice are embryonically lethal at E10.5 and show degeneration defects in late foregut development[32]. Since pancreatic tissue is posterior-foregut derived[315], and endodermal lncRNAs e.g., *DEANR1* have been implicated in successful pancreatic development[316], we wanted to functionally validate *LNC SOX17* absent and transcriptomically aberrant definitive endoderm. Hence, we carried out *in vitro* pancreatic differentiation beyond definitive endoderm[247]. Doing so we find *LNC SOX17* lacking definitive endoderm to generate less PP1 cells accompanied by overall decreased *PDX1* expression. In concordance with these data, transcriptomes of *LNC SOX17* ablated cells show a completely absent PP1 marker gene expression profile accompanied by transcriptome divergence for the most differentially expressed genes at day 9 (terminal PP1 differentiation time-point). These results are highly in concordance with the fact that lncRNAs exhibit higher tissue-specificity than protein coding-genes[295] to guarantee proper fate-decision-making in a developmental context[301]. One may speculate that, if *LNC SOX17* regulates its genome in *trans* during definitive endoderm to maintain a respective transcriptome – e.g. favoured for posteriorization of the foregut – anterior foregut derived tissue differentiations of *LNC SOX17* depleted and control cells shall be comparably effective. The hypothesis of *LNC SOX17* potentially being involved in priming gut-patterning is plausible, since under *in vitro* culture conditions cells receive unique signaling which is not the case in a developing embryo. As for instance when exposing mouse *Brachyury<sup>+</sup>/Foxa2<sup>lo</sup>* posterior primitive streak populations to high levels of Activin, these cells are still able to generate endoderm, indicating that germ-layer fates are not

yet fixed at the primitive streak stage in mESC differentiation cultures[15]. That said, *in vivo* there is potentially more space for plasticity even within human definitive endoderm, as we find only a small fraction of cells expressing *LNC SOX17* in the human gastrula[17]. This observation may either pinpoint technical obstacles and sequencing limitations or *LNC SOX17* to be expressed in an endodermal subcluster as described by the authors (DE1, DE2, hypoblast (Hypo) and yolk sac endoderm (YSE))[17]. Since the presence of *LNC SOX17* is conserved between human and mouse and its promoter *pLNC SOX17* is highly conserved among vertebrates, one could also think about genetic ablation models to test these hypotheses *in vivo*. These experiments would help to understand how dependent the formation of *in vivo* definitive endoderm on the presence of *LNC SOX17* actually is and may help to identify and understand the link between molecular and cellular phenotypes.

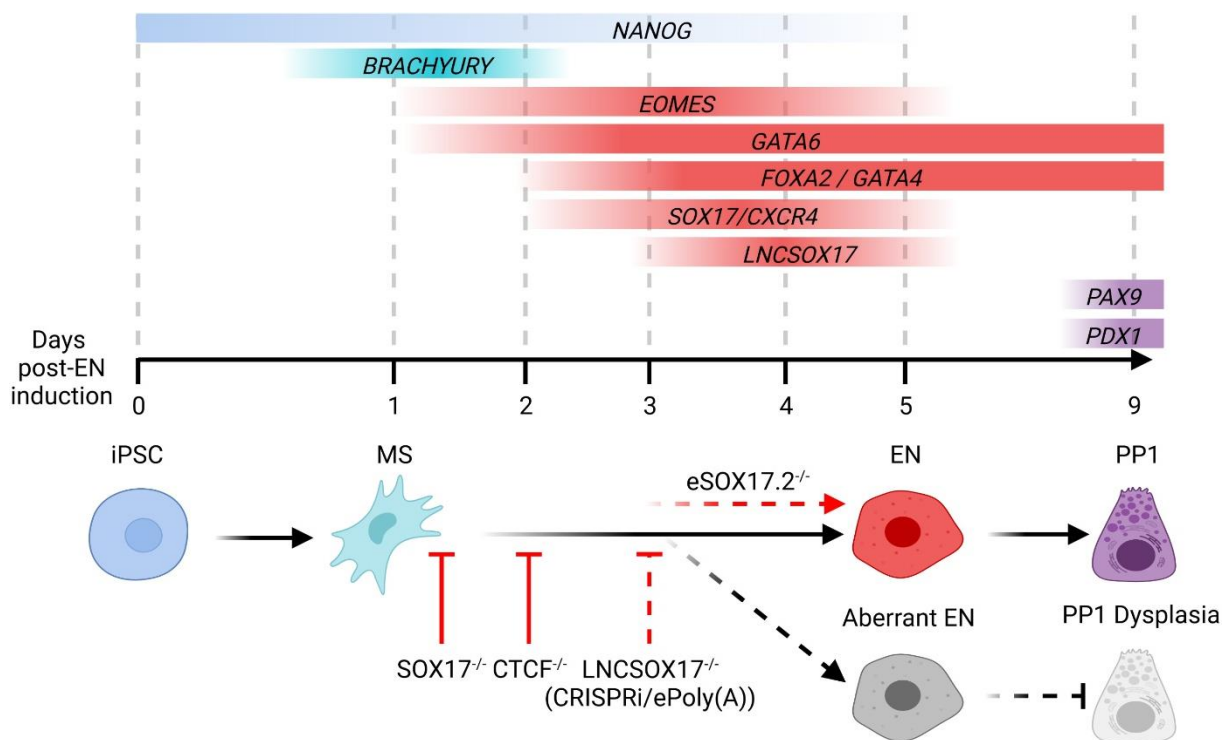
## 5.5 Conclusion

Based on both our studies, this thesis overall highlights the importance of *cis*-regulation (Fig. 34, left) at the *SOX17* locus and beyond (Fig. 34, right), during the formation of human definitive endoderm. We define key genetic and transcriptional determinants and describe their involvement and functional relevance regarding *SOX17* gene-control. Relevant for *SOX17 cis*-regulation, we identify its centromeric CTCF-boundary maintaining and safeguarding 3D chromatin architecture at the locus and its cognate distal enhancer within the *SOX17*-DMR (*eSOX17* with its core *eSOX17.2*). We further mention the presence of “shadow-enhancers” downstream of *eSOX17*, potentially involved in *SOX17 cis*-regulation. Finally, we identify a novel lncRNA locus namely *LNCXOX17* with its cognate promoter *pLNCXOX17*. *LNCXOX17s* active transcription and produced lncRNA are both found not to be involved in *cis*- (Fig. 34, left), but *LNCXOX17* potentially in *trans*-regulation (Fig. 34, right) of distinct targets within the genome, to safeguard definitive endodermal fate and assure proper development.



**Fig. 34 Cis- versus trans-regulation of *SOX17* and its genetic determinants in definitive endoderm.** Depicted in the left panel, *cis*-regulation of *SOX17* by its genetic determinants in definitive endoderm. The *SOX17* gene (purple) gets activated by its distal enhancer *eSOX17* (orange) specifically in definitive endoderm. This regulation is highly dependent on the CTCF-loop formation, facilitating proximity between *eSOX17* and *pSOX17*. Potential shadow enhancers (orange) in proximity to *eSOX17* may compensate for deregulated *eSOX17*-*pSOX17* interaction and transcription of *SOX17*. Transcribed *SOX17* mRNA (black mRNA) is translated further into TF-protein (green). *SOX17* induces expression of *LNCXOX17* (black lncRNA) by binding at its promoter *pLNCXOX17* inducing *LNCXOX17* transcription. *LNCXOX17* does not influence chromatin architecture, *eSOX17*-*pSOX17* contacts or transcription of *SOX17* mRNA in *cis*. As depicted in the right panel, *LNCXOX17* is moreover hypothesized to regulate loci in the genome in *trans* (?), potentially via the interaction of unknown proteins (?). Unclear is how (directly/indirectly) this regulation is facilitated and if *LNCXOX17* can regulate active (red arrow) as well as inactive (green arrow) loci or just a certain type. Also unclear is, if and how the observed phenotypes are linked to *LNCXOX17* facilitated regulation. (Created with BioRender.com)

Additionally, we associate different perturbations at the *SOX17* locus and their resulting alterations in gene-control to molecular and cellular phenotypes (Fig. 35). We show whole *SOX17* gene-body ablation to lack late but not early endodermal marker induction including *LNC**SOX17*, potentially exiting pluripotency without the ability to enter definitive endoderm. Further, we show that boundary-perturbation based loss of enhancer-promoter interactions at the *SOX17* locus, lead to a “mesendodermal like” phenotype highly associated with de-regulated WNT-signaling, rescued by ectopic *SOX17*. We identify core-enhancer perturbations to lead to delayed *SOX17*-induction, potentially compensated by functional “shadow enhancers”. Finally, we find *LNC**SOX17* RNA but not transcriptional ablation to create a variety of molecular phenotypes (Abberant transcriptome, EMT-failure, JNK/JUN/AP-1 hyperactivity), leading to insufficient posterior-foregut derived PP1 pancreatic progenitor differentiation.



**Fig. 35 Association of genetic dissections at the *SOX17* locus with definitive endoderm and pancreatic differentiation phenotypes.** Depicted in the upper panel, temporal expression profiles of key genes during wild type definitive endoderm and early pancreatic progenitor differentiation. Below shown are the respective cell-states (induced pluripotent stem cells = iPSC, Mesendoderm = ME, Definitive Endoderm = EN, Early Pancreatic Progenitors = PP1). *SOX17* gene-body perturbation (*SOX17*<sup>-/-</sup>) leads to a lack of *SOX17*, *LNC**SOX17*, *FOXA2* and *GATA4* expression including decreased *NANOG* levels along with induction of master TF *GATA6* in bulk EN, highlighting exit of pluripotency but EN differentiation failure. *CTCF* loop domain Boundary 2 perturbation (*CTCF*<sup>-/-</sup>) leads to a stalled “mesendodermal like” state due to massively reduced *SOX17* levels in a major fraction of differentiating cells (*CXCR4*<sup>-</sup>). The minor cell fraction (*CXCR4*<sup>+</sup>) of this perturbation model induces *SOX17* but is transcriptomically aberrant compared to wild type *SOX17* expressing cells. Enhancer perturbation (*eSOX17.2*<sup>-/-</sup>) in an intact *CTCF* loop-domain leads to a reduction of cells inducing *SOX17* on day 3 but keeping up to wild-type *SOX17* levels upon day 5. *LNC**SOX17* repression models (CRISPRi/ePoly(A)) do not affect *SOX17* expression and lead to a transcriptionally aberrant EN that fails to perform EMT, shows JNK hyperactivity and is not capable of giving rise to PP1 cells expressing proper *PDX1* levels. (Created with BioRender.com)

## 6 REFERENCES

1. Wolpert, L., *The triumph of the embryo*. 1993, Oxford; New York: Oxford University Press.
2. Carlson, B.M., *Chapter 5 - Formation of Germ Layers and Early Derivatives*, in *Human Embryology and Developmental Biology (Fifth Edition)*, B.M. Carlson, Editor. 2014, W.B. Saunders: Philadelphia. p. 75-91.
3. Ghimire, S., et al., *Human gastrulation: The embryo and its models*. *Dev Biol*, 2021. **474**: p. 100-108.
4. Brown, J.J. and V.E. Papaioannou, *Ontogeny of hyaluronan secretion during early mouse development*. *Development*, 1993. **117**(2): p. 483-92.
5. Shahbazi, M.N. and M. Zernicka-Goetz, *Deconstructing and reconstructing the mouse and human early embryo*. *Nat Cell Biol*, 2018. **20**(8): p. 878-887.
6. Shao, Y., et al., *A pluripotent stem cell-based model for post-implantation human amniotic sac development*. *Nat Commun*, 2017. **8**(1): p. 208.
7. D'Amour, K.A., et al., *Efficient differentiation of human embryonic stem cells to definitive endoderm*. *Nat Biotechnol*, 2005. **23**(12): p. 1534-41.
8. Yiangou, L., et al., *Human Pluripotent Stem Cell-Derived Endoderm for Modeling Development and Clinical Applications*. *Cell Stem Cell*, 2018. **22**(4): p. 485-499.
9. Kubo, A., et al., *Development of definitive endoderm from embryonic stem cells in culture*. *Development*, 2004. **131**(7): p. 1651-62.
10. Yasunaga, M., et al., *Induction and monitoring of definitive and visceral endoderm differentiation of mouse ES cells*. *Nat Biotechnol*, 2005. **23**(12): p. 1542-50.
11. Gouon-Evans, V., et al., *BMP-4 is required for hepatic specification of mouse embryonic stem cell-derived definitive endoderm*. *Nat Biotechnol*, 2006. **24**(11): p. 1402-11.
12. Tada, S., et al., *Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture*. *Development*, 2005. **132**(19): p. 4363-74.
13. Gadue, P., et al., *Wnt and TGF-beta signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells*. *Proc Natl Acad Sci U S A*, 2006. **103**(45): p. 16806-11.
14. Schier, A.F. and M.M. Shen, *Nodal signalling in vertebrate development*. *Nature*, 2000. **403**(6768): p. 385-9.
15. Wells, J.M. and D.A. Melton, *Vertebrate endoderm development*. *Annu Rev Cell Dev Biol*, 1999. **15**: p. 393-410.
16. McGrath, K.E., et al., *Embryonic expression and function of the chemokine SDF-1 and its receptor, CXCR4*. *Dev Biol*, 1999. **213**(2): p. 442-56.
17. Tyser, R.C.V., et al., *Single-cell transcriptomic characterization of a gastrulating human embryo*. *Nature*, 2021. **600**(7888): p. 285-289.
18. Pijuan-Sala, B., et al., *A single-cell molecular map of mouse gastrulation and early organogenesis*. *Nature*, 2019. **566**(7745): p. 490-495.
19. Grosswendt, S., et al., *Epigenetic regulator function through mouse gastrulation*. *Nature*, 2020. **584**(7819): p. 102-108.
20. Warmflash, A., et al., *A method to recapitulate early embryonic spatial patterning in human embryonic stem cells*. *Nat Methods*, 2014. **11**(8): p. 847-54.

21. Gifford, C.A., et al., *Transcriptional and epigenetic dynamics during specification of human embryonic stem cells*. Cell, 2013. **153**(5): p. 1149-63.
22. Sozen, B., et al., *Self-assembly of embryonic and two extra-embryonic stem cell types into gastrulating embryo-like structures*. Nat Cell Biol, 2018. **20**(8): p. 979-989.
23. Beccari, L., et al., *Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids*. Nature, 2018. **562**(7726): p. 272-276.
24. Ng, E.S., et al., *The primitive streak gene *Mixl1* is required for efficient haematopoiesis and BMP4-induced ventral mesoderm patterning in differentiating ES cells*. Development, 2005. **132**(5): p. 873-84.
25. Wiles, M.V. and B.M. Johansson, *Embryonic Stem Cell Development in a Chemically Defined Medium*. Experimental Cell Research, 1999. **247**(1): p. 241-248.
26. Itskovitz-Eldor, J., et al., *Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers*. Mol Med, 2000. **6**(2): p. 88-95.
27. Pour, M., et al., *Emergence and patterning dynamics of mouse-definitive endoderm*. iScience, 2022. **25**(1): p. 103556.
28. Choi, J., et al., *A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs*. Nat Biotechnol, 2015. **33**(11): p. 1173-81.
29. Tsankov, A.M., et al., *Transcription factor binding dynamics during human ES cell differentiation*. Nature, 2015. **518**(7539): p. 344-9.
30. Anani, M., I. Nobuhisa, and T. Taga, *Sry-related High Mobility Group Box 17 Functions as a Tumor Suppressor by Antagonizing the Wingless-related Integration Site Pathway*. Journal of Cancer Prevention, 2020. **25**(4): p. 204-212.
31. Irie, N., et al., *SOX17 is a critical specifier of human primordial germ cell fate*. Cell, 2015. **160**(1-2): p. 253-68.
32. Kanai-Azuma, M., et al., *Depletion of definitive gut endoderm in Sox17-null mutant mice*. Development, 2002. **129**(10): p. 2367-79.
33. Mukherjee, S., et al., *Sox17 and beta-catenin co-occupy Wnt-responsive enhancers to govern the endoderm gene regulatory network*. Elife, 2020. **9**.
34. Shimoda, M., et al., *Sox17 plays a substantial role in late-stage differentiation of the extraembryonic endoderm in vitro*. J Cell Sci, 2007. **120**(Pt 21): p. 3859-69.
35. Sinner, D., et al., *Sox17 and beta-catenin cooperate to regulate the transcription of endodermal genes*. Development, 2004. **131**(13): p. 3069-80.
36. Bowles, J., G. Schepers, and P. Koopman, *Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators*. Dev Biol, 2000. **227**(2): p. 239-55.
37. Gubbay, J., et al., *A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes*. Nature, 1990. **346**(6281): p. 245-50.
38. Jay, P., et al., *SOX22 is a new member of the SOX gene family, mainly expressed in human nervous tissue*. Hum Mol Genet, 1997. **6**(7): p. 1069-77.
39. Kamachi, Y., et al., *Involvement of Sox1, 2 and 3 in the early and subsequent molecular events of lens induction*. Development, 1998. **125**(13): p. 2521-32.
40. Lefebvre, V., P. Li, and B. de Crombrughe, *A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene*. EMBO J, 1998. **17**(19): p. 5718-33.

41. Uchikawa, M., Y. Kamachi, and H. Kondoh, *Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken*. *Mech Dev*, 1999. **84**(1-2): p. 103-20.
42. Schepers, G.E., et al., *Cloning and characterisation of the Sry-related transcription factor gene Sox8*. *Nucleic Acids Res*, 2000. **28**(6): p. 1473-80.
43. Jimenez, G., et al., *Relief of gene repression by torso RTK signaling: role of capicua in Drosophila terminal and dorsoventral patterning*. *Genes Dev*, 2000. **14**(2): p. 224-31.
44. Ohno, S., *Evolution by Gene Duplication*. 1970, Springer Berlin Heidelberg: Berlin, Heidelberg.
45. Patthy, L., *Modular exchange principles in proteins*. *Current Opinion in Structural Biology*, 1991. **1**(3): p. 351-361.
46. Patthy, L., *Introns and exons*. *Current Opinion in Structural Biology*, 1994. **4**(3): p. 383-392.
47. Holland, P.W., et al., *Gene duplications and the origins of vertebrate development*. *Dev Suppl*, 1994: p. 125-33.
48. Wegner, M., *From head to toes: the multiple facets of Sox proteins*. *Nucleic Acids Res*, 1999. **27**(6): p. 1409-20.
49. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations*. *Genetics*, 1999. **151**(4): p. 1531-45.
50. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization*. *Genetics*, 2000. **154**(1): p. 459-73.
51. MacCarthy, T. and A. Bergman, *The limits of subfunctionalization*. *BMC Evol Biol*, 2007. **7**: p. 213.
52. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. *Nature*, 2001. **409**(6822): p. 860-921.
53. Alexander, R.P., et al., *Annotating non-coding regions of the genome*. *Nat Rev Genet*, 2010. **11**(8): p. 559-71.
54. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome*. *Nature*, 2012. **489**(7414): p. 57-74.
55. Morris, K.V. and J.S. Mattick, *The rise of regulatory RNA*. *Nat Rev Genet*, 2014. **15**(6): p. 423-37.
56. Levine, M. and R. Tjian, *Transcription regulation and animal diversity*. *Nature*, 2003. **424**(6945): p. 147-51.
57. Gaszner, M. and G. Felsenfeld, *Insulators: exploiting transcriptional and epigenetic mechanisms*. *Nat Rev Genet*, 2006. **7**(9): p. 703-13.
58. Maeda, R.K. and F. Karch, *Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions*. *Current Opinion in Genetics & Development*, 2011. **21**(2): p. 187-193.
59. Bell, A.C., A.G. West, and G. Felsenfeld, *The protein CTCF is required for the enhancer blocking activity of vertebrate insulators*. *Cell*, 1999. **98**(3): p. 387-96.
60. Klenova, E.M., et al., *Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF*. *Mol Cell Biol*, 2001. **21**(6): p. 2221-34.
61. Arnone, M.I. and E.H. Davidson, *The hardwiring of development: organization and function of genomic regulatory systems*. *Development*, 1997. **124**(10): p. 1851-64.



62. Haberle, V. and A. Stark, *Eukaryotic core promoters and the functional basis of transcription initiation*. Nat Rev Mol Cell Biol, 2018. **19**(10): p. 621-637.
63. Hampsey, M., *Molecular genetics of the RNA polymerase II general transcriptional machinery*. Microbiol Mol Biol Rev, 1998. **62**(2): p. 465-503.
64. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription*. Genes Dev, 2011. **25**(10): p. 1010-22.
65. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters*. Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1412-7.
66. Daneshvar, K., et al., *DIGIT Is a Conserved Long Noncoding RNA that Regulates GSC Expression to Control Definitive Endoderm Differentiation of Embryonic Stem Cells*. Cell Rep, 2016. **17**(2): p. 353-365.
67. Mikhaylichenko, O., et al., *The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription*. Genes Dev, 2018. **32**(1): p. 42-57.
68. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences*. Cell, 1981. **27**(2 Pt 1): p. 299-308.
69. Moreau, P., et al., *The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants*. Nucleic Acids Res, 1981. **9**(22): p. 6047-68.
70. Tuan, D., S. Kong, and K. Hu, *Transcription of the hypersensitive site HS2 enhancer in erythroid cells*. Proc Natl Acad Sci U S A, 1992. **89**(23): p. 11219-23.
71. Koch, F., et al., *Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters*. Nat Struct Mol Biol, 2011. **18**(8): p. 956-63.
72. Bulger, M. and M. Groudine, *Functional and mechanistic diversity of distal transcription enhancers*. Cell, 2011. **144**(3): p. 327-39.
73. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly*. Hum Mol Genet, 2003. **12**(14): p. 1725-35.
74. Long, H.K., et al., *Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder*. Cell Stem Cell, 2020. **27**(5): p. 765-783 e14.
75. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture*. Nature, 2010. **467**(7314): p. 430-5.
76. Furlong, E.E.M. and M. Levine, *Developmental enhancers and chromosome topology*. Science, 2018. **361**(6409): p. 1341-1345.
77. Mitchell, P.J. and R. Tjian, *Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins*. Science, 1989. **245**(4916): p. 371-8.
78. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control*. Nat Rev Genet, 2012. **13**(9): p. 613-26.
79. Malik, S. and R.G. Roeder, *Dynamic regulation of pol II transcription by the mammalian Mediator complex*. Trends Biochem Sci, 2005. **30**(5): p. 256-63.
80. Cho, W.K., et al., *Mediator and RNA polymerase II clusters associate in transcription-dependent condensates*. Science, 2018. **361**(6400): p. 412-415.
81. Hnisz, D., et al., *A Phase Separation Model for Transcriptional Control*. Cell, 2017. **169**(1): p. 13-23.
82. Sabari, B.R., et al., *Coactivator condensation at super-enhancers links phase separation and gene control*. Science, 2018. **361**(6400).

83. Boija, A., et al., *Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains*. Cell, 2018. **175**(7): p. 1842-1855 e16.
84. Basu, S., et al., *Unblending of Transcriptional Condensates in Human Repeat Expansion Disease*. Cell, 2020. **181**(5): p. 1062-1079 e30.
85. Shrinivas, K., et al., *Enhancer Features that Drive Formation of Transcriptional Condensates*. Mol Cell, 2019. **75**(3): p. 549-561 e7.
86. Henninger, J.E., et al., *RNA-Mediated Feedback Control of Transcriptional Condensates*. Cell, 2021. **184**(1): p. 207-225 e24.
87. Chen, H., et al., *Dynamic interplay between enhancer-promoter topology and gene activity*. Nat Genet, 2018. **50**(9): p. 1296-1303.
88. Deng, W., et al., *Reactivation of developmentally silenced globin genes by forced chromatin looping*. Cell, 2014. **158**(4): p. 849-860.
89. Paliou, C., et al., *Prefomed chromatin topology assists transcriptional robustness of Shh during limb development*. Proc Natl Acad Sci U S A, 2019. **116**(25): p. 12390-12399.
90. Williamson, I., et al., *Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity*. Development, 2016. **143**(16): p. 2994-3001.
91. Wu, H.J., et al., *Topological isolation of developmental regulators in mammalian genomes*. Nat Commun, 2021. **12**(1): p. 4897.
92. Huang, Y., R. Neijts, and W. de Laat, *How chromosome topologies get their shape: views from proximity ligation and microscopy methods*. FEBS Lett, 2020. **594**(21): p. 3439-3449.
93. Lichter, P., et al., *Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries*. Hum Genet, 1988. **80**(3): p. 224-34.
94. Trask, B., D. Pinkel, and G. van den Engh, *The proximity of DNA sequences in interphase cell nuclei is correlated to genomic distance and permits ordering of cosmids spanning 250 kilobase pairs*. Genomics, 1989. **5**(4): p. 710-7.
95. Cremer, T. and C. Cremer, *Chromosome territories, nuclear architecture and gene regulation in mammalian cells*. Nat Rev Genet, 2001. **2**(4): p. 292-301.
96. Haaf, T. and M. Schmid, *Chromosome topology in mammalian interphase nuclei*. Exp Cell Res, 1991. **192**(2): p. 325-32.
97. Dekker, J., et al., *Capturing chromosome conformation*. Science, 2002. **295**(5558): p. 1306-11.
98. Dekker, J., M.A. Marti-Renom, and L.A. Mirny, *Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data*. Nat Rev Genet, 2013. **14**(6): p. 390-403.
99. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*. Science, 2009. **326**(5950): p. 289-93.
100. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions*. Nature, 2012. **485**(7398): p. 376-80.
101. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre*. Nature, 2012. **485**(7398): p. 381-5.
102. Ruf, S., et al., *Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor*. Nat Genet, 2011. **43**(4): p. 379-86.

103. Symmons, O., et al., *Functional and topological characteristics of mammalian regulatory domains*. *Genome Res*, 2014. **24**(3): p. 390-400.
104. Hnisz, D., D.S. Day, and R.A. Young, *Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control*. *Cell*, 2016. **167**(5): p. 1188-1200.
105. Chepelev, I., et al., *Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization*. *Cell Res*, 2012. **22**(3): p. 490-503.
106. DeMare, L.E., et al., *The genomic landscape of cohesin-associated chromatin interactions*. *Genome Res*, 2013. **23**(8): p. 1224-34.
107. Downen, J.M., et al., *Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes*. *Cell*, 2014. **159**(2): p. 374-387.
108. Nasmyth, K. and C.H. Haering, *Cohesin: its roles and mechanisms*. *Annu Rev Genet*, 2009. **43**: p. 525-58.
109. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping*. *Cell*, 2014. **159**(7): p. 1665-80.
110. Fullwood, M.J., et al., *An oestrogen-receptor-alpha-bound human chromatin interactome*. *Nature*, 2009. **462**(7269): p. 58-64.
111. Handoko, L., et al., *CTCF-mediated functional chromatin interactome in pluripotent cells*. *Nat Genet*, 2011. **43**(7): p. 630-8.
112. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment*. *Cell*, 2013. **153**(6): p. 1281-95.
113. Tang, Z., et al., *CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription*. *Cell*, 2015. **163**(7): p. 1611-27.
114. Fudenberg, G., et al., *Formation of Chromosomal Domains by Loop Extrusion*. *Cell Rep*, 2016. **15**(9): p. 2038-49.
115. Nuebler, J., et al., *Chromatin organization by an interplay of loop extrusion and compartmental segregation*. *Proc Natl Acad Sci U S A*, 2018. **115**(29): p. E6697-E6706.
116. de Wit, E., et al., *CTCF Binding Polarity Determines Chromatin Looping*. *Mol Cell*, 2015. **60**(4): p. 676-84.
117. Guo, Y., et al., *CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function*. *Cell*, 2015. **162**(4): p. 900-10.
118. Sanborn, A.L., et al., *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes*. *Proc Natl Acad Sci U S A*, 2015. **112**(47): p. E6456-65.
119. Busslinger, G.A., et al., *Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl*. *Nature*, 2017. **544**(7651): p. 503-507.
120. Haarhuis, J.H.I., et al., *The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension*. *Cell*, 2017. **169**(4): p. 693-707 e14.
121. Wutz, G., et al., *Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins*. *EMBO J*, 2017. **36**(24): p. 3573-3599.
122. Rao, S.S.P., et al., *Cohesin Loss Eliminates All Loop Domains*. *Cell*, 2017. **171**(2): p. 305-320 e24.
123. Nora, E.P., et al., *Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization*. *Cell*, 2017. **169**(5): p. 930-944 e22.

124. Schwarzer, W., et al., *Two independent modes of chromatin organization revealed by cohesin removal*. Nature, 2017. **551**(7678): p. 51-56.
125. de Laat, W. and D. Duboule, *Topology of mammalian developmental enhancers and their regulatory landscapes*. Nature, 2013. **502**(7472): p. 499-506.
126. Bintu, B., et al., *Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells*. Science, 2018. **362**(6413).
127. Stevens, T.J., et al., *3D structures of individual mammalian genomes studied by single-cell Hi-C*. Nature, 2017. **544**(7648): p. 59-64.
128. Szabo, Q., et al., *Regulation of single-cell genome organization into TADs and chromatin nanodomains*. Nat Genet, 2020. **52**(11): p. 1151-1157.
129. Bonev, B., et al., *Multiscale 3D Genome Rewiring during Mouse Neural Development*. Cell, 2017. **171**(3): p. 557-572 e24.
130. Stik, G., et al., *CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response*. Nature Genetics, 2020. **52**(7): p. 655-661.
131. Hsieh, C.H., et al., *Functional Impairment in Mito Degradation and Mitophagy Is a Shared Feature in Familial and Sporadic Parkinson's Disease*. Cell Stem Cell, 2016. **19**(6): p. 709-724.
132. Krietenstein, N., et al., *Ultrastructural Details of Mammalian Chromosome Architecture*. Mol Cell, 2020. **78**(3): p. 554-565 e7.
133. Mumbach, M.R., et al., *HiChIP: efficient and sensitive analysis of protein-directed genome architecture*. Nat Methods, 2016. **13**(11): p. 919-922.
134. Andrey, G., et al., *Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding*. Genome Res, 2017. **27**(2): p. 223-233.
135. Kubo, N., et al., *Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation*. Nat Struct Mol Biol, 2021. **28**(2): p. 152-161.
136. Hansen, A.S., et al., *Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF*. Mol Cell, 2019. **76**(3): p. 395-411 e13.
137. Flavahan, W.A., et al., *Insulator dysfunction and oncogene activation in IDH mutant gliomas*. Nature, 2016. **529**(7584): p. 110-4.
138. Frith, M.C., M. Pheasant, and J.S. Mattick, *The amazing complexity of the human transcriptome*. Eur J Hum Genet, 2005. **13**(8): p. 894-7.
139. Kapranov, P., A.T. Willingham, and T.R. Gingeras, *Genome-wide transcription and the implications for genomic organization*. Nat Rev Genet, 2007. **8**(6): p. 413-23.
140. Mattick, J.S. and I.V. Makunin, *Non-coding RNA*. Hum Mol Genet, 2006. **15 Spec No 1**: p. R17-29.
141. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature, 2007. **447**(7146): p. 799-816.
142. Carninci, P., et al., *The transcriptional landscape of the mammalian genome*. Science, 2005. **309**(5740): p. 1559-63.
143. Kapranov, P., et al., *Large-scale transcriptional activity in chromosomes 21 and 22*. Science, 2002. **296**(5569): p. 916-9.
144. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays*. Science, 2004. **306**(5705): p. 2242-6.
145. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome*. Nature, 2004. **431**(7011): p. 931-45.

146. Riddihough, G., *In the Forests of RNA Dark Matter*. Science, 2005. **309**(5740): p. 1507-1507.
147. Johnson, J.M., et al., *Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments*. Trends Genet, 2005. **21**(2): p. 93-102.
148. Statello, L., et al., *Gene regulation by long non-coding RNAs and its biological functions*. Nat Rev Mol Cell Biol, 2021. **22**(2): p. 96-118.
149. Lam, M.T., et al., *Enhancer RNAs and regulated transcriptional programs*. Trends Biochem Sci, 2014. **39**(4): p. 170-82.
150. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers*. PLoS Biol, 2010. **8**(5): p. e1000384.
151. Core, L.J., et al., *Defining the status of RNA polymerase at promoters*. Cell Rep, 2012. **2**(4): p. 1025-35.
152. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues*. Nature, 2014. **507**(7493): p. 455-461.
153. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers*. Nature, 2010. **465**(7295): p. 182-7.
154. Melgar, M.F., F.S. Collins, and P. Sethupathy, *Discovery of active enhancers through bidirectional expression of short transcripts*. Genome Biol, 2011. **12**(11): p. R113.
155. Rank, G., M. Prestel, and R. Paro, *Transcription through intergenic chromosomal memory elements of the Drosophila bithorax complex correlates with an epigenetic switch*. Mol Cell Biol, 2002. **22**(22): p. 8026-34.
156. Schmitt, S., M. Prestel, and R. Paro, *Intergenic transcription through a polycomb group response element counteracts silencing*. Genes Dev, 2005. **19**(6): p. 697-708.
157. Paralkar, V.R., et al., *Unlinking an lncRNA from Its Associated cis Element*. Mol Cell, 2016. **62**(1): p. 104-10.
158. Li, M., et al., *A putative long noncoding RNA-encoded micropeptide maintains cellular homeostasis in pancreatic beta cells*. Mol Ther Nucleic Acids, 2021. **26**: p. 307-320.
159. van Heesch, S., et al., *The Translational Landscape of the Human Heart*. Cell, 2019. **178**(1): p. 242-260 e29.
160. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression*. Genome Res, 2012. **22**(9): p. 1775-89.
161. Uszczyńska-Ratajczak, B., et al., *Towards a complete map of the human long non-coding RNA transcriptome*. Nat Rev Genet, 2018. **19**(9): p. 535-548.
162. Fang, S., et al., *NONCODEV5: a comprehensive annotation database for long non-coding RNAs*. Nucleic Acids Res, 2018. **46**(D1): p. D308-D314.
163. Mele, M., et al., *Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs*. Genome Res, 2017. **27**(1): p. 27-37.
164. Mattioli, K., et al., *High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity*. Genome Res, 2019. **29**(3): p. 344-355.
165. Andergassen, D. and J.L. Rinn, *From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo*. Nat Rev Genet, 2021.
166. Hung, T. and H.Y. Chang, *Long noncoding RNA in genome regulation: prospects and mechanisms*. RNA Biol, 2010. **7**(5): p. 582-5.
167. Bonasio, R., S. Tu, and D. Reinberg, *Molecular signals of epigenetic states*. Science, 2010. **330**(6004): p. 612-6.

168. Wang, K.C. and H.Y. Chang, *Molecular mechanisms of long noncoding RNAs*. Mol Cell, 2011. **43**(6): p. 904-14.
169. Hung, T., et al., *Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters*. Nat Genet, 2011. **43**(7): p. 621-9.
170. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor*. Sci Signal, 2010. **3**(107): p. ra8.
171. Spitale, R.C., M.C. Tsai, and H.Y. Chang, *RNA templating the epigenome: long noncoding RNAs as molecular scaffolds*. Epigenetics, 2011. **6**(5): p. 539-43.
172. Huarte, M., et al., *A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response*. Cell, 2010. **142**(3): p. 409-19.
173. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis*. Nature, 2010. **464**(7291): p. 1071-6.
174. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs*. Cell, 2007. **129**(7): p. 1311-23.
175. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells*. Cell, 2010. **143**(1): p. 46-58.
176. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation*. Nature, 2011. **477**(7364): p. 295-300.
177. Wang, P., et al., *Targeting SOX17 in human embryonic stem cells creates unique strategies for isolating and analyzing developing endoderm*. Cell Stem Cell, 2011. **8**(3): p. 335-46.
178. Chu, L.F., et al., *Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm*. Genome Biol, 2016. **17**(1): p. 173.
179. Chia, C.Y., et al., *GATA6 Cooperates with EOMES/SMAD2/3 to Deploy the Gene Regulatory Network Governing Human Definitive Endoderm and Pancreas Formation*. Stem Cell Reports, 2019. **12**(1): p. 57-70.
180. Burtscher, I., et al., *The Sox17-mCherry fusion mouse line allows visualization of endoderm and vascular endothelial development*. Genesis, 2012. **50**(6): p. 496-505.
181. Engert, S., et al., *Sox17-2A-iCre: a knock-in mouse line expressing Cre recombinase in endoderm and vascular endothelial cells*. Genesis, 2009. **47**(9): p. 603-10.
182. Liao, W.P., et al., *Generation of a mouse line expressing Sox17-driven Cre recombinase with specific activity in arteries*. Genesis, 2009. **47**(7): p. 476-83.
183. Reddy, J., et al., *Predicting master transcription factors from pan-cancer expression data*. Sci Adv, 2021. **7**(48): p. eabf6123.
184. Hotchkiss, R.D., *The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography*. J Biol Chem, 1948. **175**(1): p. 315-32.
185. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development*. Science, 1975. **187**(4173): p. 226-32.
186. Chen, Z.X. and A.D. Riggs, *DNA methylation and demethylation in mammals*. J Biol Chem, 2011. **286**(21): p. 18347-53.
187. Haggerty, C., et al., *Dnmt1 has de novo activity targeted to transposable elements*. Nat Struct Mol Biol, 2021. **28**(7): p. 594-603.
188. Hermann, A., R. Goyal, and A. Jeltsch, *The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites*. J Biol Chem, 2004. **279**(46): p. 48350-9.

189. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development*. Cell, 1999. **99**(3): p. 247-57.
190. Suzuki, M.M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics*. Nat Rev Genet, 2008. **9**(6): p. 465-76.
191. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development*. Nat Rev Genet, 2013. **14**(3): p. 204-20.
192. Karlsson, M., et al., *A single-cell type transcriptomics map of human tissues*. Sci Adv, 2021. **7**(31).
193. Seguin, C.A., et al., *Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells*. Cell Stem Cell, 2008. **3**(2): p. 182-95.
194. Zhang, Y., et al., *SOX17 is a tumor suppressor in endometrial cancer*. Oncotarget, 2016. **7**(46).
195. Merino-Azpitarte, M., et al., *SOX17 regulates cholangiocyte differentiation and acts as a tumor suppressor in cholangiocarcinoma*. J Hepatol, 2017. **67**(1): p. 72-83.
196. Zhan, T., N. Rindtorff, and M. Boutros, *Wnt signaling in cancer*. Oncogene, 2017. **36**(11): p. 1461-1473.
197. Li, L., et al., *SOX17 restrains proliferation and tumor formation by down-regulating activity of the Wnt/beta-catenin signaling pathway via trans-suppressing beta-catenin in cervical cancer*. Cell Death Dis, 2018. **9**(7): p. 741.
198. Zhou, W., et al., *SOX17 Inhibits Tumor Metastasis Via Wnt Signaling In Endometrial Cancer*. Onco Targets Ther, 2019. **12**: p. 8275-8286.
199. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts*. Genome Res, 2014. **24**(6): p. 999-1011.
200. Rowley, M.J., et al., *Analysis of Hi-C data using SIP effectively identifies loops in organisms from C. elegans to mammals*. Genome Res, 2020. **30**(3): p. 447-458.
201. Durand, N.C., et al., *Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments*. Cell Syst, 2016. **3**(1): p. 95-8.
202. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities*. Mol Cell, 2010. **38**(4): p. 576-89.
203. Ji, X., et al., *3D Chromosome Regulatory Landscape of Human Pluripotent Cells*. Cell Stem Cell, 2016. **18**(2): p. 262-75.
204. Chen, X., et al., *Key role for CTCF in establishing chromatin structure in human embryos*. Nature, 2019. **576**(7786): p. 306-310.
205. Lee, M.T., A.R. Bonneau, and A.J. Giraldez, *Zygotic genome activation during the maternal-to-zygotic transition*. Annu Rev Cell Dev Biol, 2014. **30**: p. 581-613.
206. Zheng, H. and W. Xie, *The role of 3D genome organization in development and cell differentiation*. Nat Rev Mol Cell Biol, 2019. **20**(9): p. 535-550.
207. Golbus, M.S., P.G. Calarco, and C.J. Epstein, *The effects of inhibitors of RNA synthesis (alpha-amanitin and actinomycin D) on preimplantation mouse embryogenesis*. J Exp Zool, 1973. **186**(2): p. 207-16.
208. Hamatani, T., et al., *Dynamics of global gene expression changes during mouse preimplantation development*. Dev Cell, 2004. **6**(1): p. 117-31.
209. Newport, J. and M. Kirschner, *A major developmental transition in early Xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage*. Cell, 1982. **30**(3): p. 675-86.

210. Uuskula-Reimand, L., et al., *Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders*. *Genome Biol*, 2016. **17**(1): p. 182.
211. Tandon, R., et al., *Generation of two human isogenic iPSC lines from fetal dermal fibroblasts*. *Stem Cell Res*, 2018. **33**: p. 120-124.
212. Lu, L., et al., *Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases*. *Mol Cell*, 2020. **79**(3): p. 521-534 e15.
213. Pauklin, S. and L. Vallier, *The cell-cycle state of stem cells determines cell fate propensity*. *Cell*, 2013. **155**(1): p. 135-47.
214. Mao, B. and C. Niehrs, *Kremen2 modulates Dickkopf2 activity during Wnt/LRP6 signaling*. *Gene*, 2003. **302**(1-2): p. 179-83.
215. Engert, S., et al., *Wnt/beta-catenin signalling regulates Sox17 expression and is essential for organizer and endoderm formation in the mouse*. *Development*, 2013. **140**(15): p. 3128-38.
216. Banaszynski, L.A., et al., *A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules*. *Cell*, 2006. **126**(5): p. 995-1004.
217. Li, Q.V., et al., *Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation*. *Nat Genet*, 2019. **51**(6): p. 999-1010.
218. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization*. *Nat Methods*, 2012. **9**(3): p. 215-6.
219. Ernst, J. and M. Kellis, *Chromatin-state discovery and genome annotation with ChromHMM*. *Nat Protoc*, 2017. **12**(12): p. 2478-2492.
220. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome*. *Science*, 2015. **347**(6220): p. 1260419.
221. Thul, P.J., et al., *A subcellular map of the human proteome*. *Science*, 2017. **356**(6340).
222. Mukherjee, N., et al., *Integrative classification of human coding and noncoding genes through RNA metabolism profiles*. *Nat Struct Mol Biol*, 2017. **24**(1): p. 86-96.
223. Lagarde, J., et al., *High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing*. *Nat Genet*, 2017. **49**(12): p. 1731-1740.
224. Schlackow, M., et al., *Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs*. *Mol Cell*, 2017. **65**(1): p. 25-38.
225. Struhl, K., *Transcriptional noise and the fidelity of initiation by RNA polymerase II*. *Nat Struct Mol Biol*, 2007. **14**(2): p. 103-5.
226. Beck, Z.T., Z. Xing, and E.J. Tran, *LncRNAs: Bridging environmental sensing and gene expression*. *RNA Biol*, 2016. **13**(12): p. 1189-1196.
227. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions*. *Bioinformatics*, 2011. **27**(13): p. i275-82.
228. Yeo, N.C., et al., *An enhanced CRISPR repressor for targeted mammalian gene regulation*. *Nat Methods*, 2018. **15**(8): p. 611-616.
229. Gilbert, L.A., et al., *Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation*. *Cell*, 2014. **159**(3): p. 647-61.
230. Allou, L., et al., *Non-coding deletions identify Maenli lincRNA as a limb-specific En1 regulator*. *Nature*, 2021. **592**(7852): p. 93-98.



231. Daneshvar, K., et al., *lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation*. Nat Cell Biol, 2020. **22**(10): p. 1211-1222.
232. Martinez Barbera, J.P., et al., *The homeobox gene Hex is required in definitive endodermal tissues for normal forebrain, liver and thyroid formation*. Development, 2000. **127**(11): p. 2433-45.
233. Grapin-Botton, A. and D. Constam, *Evolution of the mechanisms and molecular control of endoderm formation*. Mech Dev, 2007. **124**(4): p. 253-78.
234. McLean, A.B., et al., *Activin A Efficiently Specifies Definitive Endoderm from Human Embryonic Stem Cells Only When Phosphatidylinositol 3-Kinase Signaling Is Suppressed*. 2007. **25**(1): p. 29-38.
235. Teo, A.K., et al., *Pluripotency factors regulate definitive endoderm specification through eomesodermin*. Genes Dev, 2011. **25**(3): p. 238-50.
236. Aksoy, I., et al., *Klf4 and Klf5 differentially inhibit mesoderm and endoderm differentiation in embryonic stem cells*. Nat Commun, 2014. **5**: p. 3719.
237. Dettmer, R., et al., *FGF2 Inhibits Early Pancreatic Lineage Specification during Differentiation of Human Embryonic Stem Cells*. Cells, 2020. **9**(9).
238. Briggs, J., et al., *Transcriptional upregulation of SPARC, in response to c-Jun overexpression, contributes to increased motility and invasion of MCF7 breast cancer cells*. Oncogene, 2002. **21**(46): p. 7077-91.
239. Schummer, P., et al., *Specific c-Jun target genes in malignant melanoma*. Cancer Biol Ther, 2016. **17**(5): p. 486-97.
240. Florin, L., et al., *Identification of novel AP-1 target genes in fibroblasts regulated during cutaneous wound healing*. Oncogene, 2004. **23**(42): p. 7005-17.
241. van Dam, H. and M. Castellazzi, *Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis*. Oncogene, 2001. **20**(19): p. 2453-64.
242. Hoffmann, E., et al., *Transcriptional regulation of EGR-1 by the interleukin-1-JNK-MKK7-c-Jun pathway*. J Biol Chem, 2008. **283**(18): p. 12120-8.
243. Kockel, L., J.G. Homsy, and D. Bohmann, *Drosophila AP-1: lessons from an invertebrate*. Oncogene, 2001. **20**(19): p. 2347-64.
244. Raivich, G. and A. Behrens, *Role of the AP-1 transcription factor c-Jun in developing, adult and injured brain*. Prog Neurobiol, 2006. **78**(6): p. 347-63.
245. Muniyappa, H. and K.C. Das, *Activation of c-Jun N-terminal kinase (JNK) by widely used specific p38 MAPK inhibitors SB202190 and SB203580: a MLK-3-MKK7-dependent mechanism*. Cell Signal, 2008. **20**(4): p. 675-83.
246. Li, Q., et al., *A sequential EMT-MET mechanism drives the differentiation of human embryonic stem cells towards hepatocytes*. Nat Commun, 2017. **8**: p. 15166.
247. Alvarez-Dominguez, J.R., et al., *Circadian Entrainment Triggers Maturation of Human In Vitro Islets*. Cell Stem Cell, 2020. **26**(1): p. 108-122 e10.
248. Williamson, I., et al., *Developmentally regulated Shh expression is robust to TAD perturbations*. Development, 2019. **146**(19).
249. Volders, P.J., et al., *LNCipedia 5: towards a reference set of human long non-coding RNAs*. Nucleic Acids Res, 2019. **47**(D1): p. D135-D139.
250. Heidari, N., et al., *Genome-wide map of regulatory interactions in the human genome*. Genome Res, 2014. **24**(12): p. 1905-17.

251. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation*. Nature, 2015. **518**(7539): p. 331-6.
252. Won, H., et al., *Chromosome conformation elucidates regulatory relationships in developing human brain*. Nature, 2016. **538**(7626): p. 523-527.
253. Jin, F., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells*. Nature, 2013. **503**(7475): p. 290-4.
254. de Wit, E., et al., *The pluripotent genome in three dimensions is shaped around pluripotency factors*. Nature, 2013. **501**(7466): p. 227-31.
255. Downen, J.M., et al., *Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes*. Cell, 2014. **159**(2): p. 374-87.
256. Hnisz, D., et al., *Activation of proto-oncogenes by disruption of chromosome neighborhoods*. Science, 2016. **351**(6280): p. 1454-1458.
257. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions*. Cell, 2015. **161**(5): p. 1012-1025.
258. Ringel, A.R., et al., *Promoter repression and 3D-restructuring resolves divergent developmental gene expression in TADs*. bioRxiv, 2021: p. 2021.10.08.463672.
259. Ke, Y., et al., *3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis*. Cell, 2017. **170**(2): p. 367-381 e20.
260. Du, Z., et al., *Allelic reprogramming of 3D chromatin architecture during early mammalian development*. Nature, 2017. **547**(7662): p. 232-235.
261. Hug, C.B., et al., *Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription*. Cell, 2017. **169**(2): p. 216-228 e19.
262. Wike, C.L., et al., *Chromatin architecture transitions from zebrafish sperm through early embryogenesis*. Genome Res, 2021. **31**(6): p. 981-994.
263. Nakamura, R., et al., *CTCF looping is established during gastrulation in medaka embryos*. Genome Res, 2021. **31**(6): p. 968-980.
264. Niu, L., et al., *Three-dimensional folding dynamics of the *Xenopus tropicalis* genome*. Nat Genet, 2021. **53**(7): p. 1075-1087.
265. Hernandez-Hernandez, A., et al., *CTCF contributes in a critical way to spermatogenesis and male fertility*. Sci Rep, 2016. **6**: p. 28355.
266. Moore, J.M., et al., *Loss of maternal CTCF is associated with peri-implantation lethality of *Ctcf* null embryos*. PLoS One, 2012. **7**(4): p. e34915.
267. Andreu, M.J., et al., *Establishment of 3D chromatin structure after fertilization and the metabolic switch at the morula-to-blastocyst transition require CTCF*. bioRxiv, 2021.
268. Weintraub, A.S., et al., *YY1 Is a Structural Regulator of Enhancer-Promoter Loops*. Cell, 2017. **171**(7): p. 1573-1588 e28.
269. Long, H.S., et al., *Making sense of the linear genome, gene function and TADs*. Epigenetics Chromatin, 2022. **15**(1): p. 4.
270. Muro, E.M., J. Ibn-Salem, and M.A. Andrade-Navarro, *The distributions of protein coding genes within chromatin domains in relation to human disease*. Epigenetics Chromatin, 2019. **12**(1): p. 72.
271. Franke, M., et al., *Duplications disrupt chromatin architecture and rewire GPR101-enhancer communication in X-linked acrogigantism*. Am J Hum Genet, 2022. **109**(4): p. 553-570.

272. van Arensbergen, J., B. van Steensel, and H.J. Bussemaker, *In search of the determinants of enhancer-promoter interaction specificity*. Trends Cell Biol, 2014. **24**(11): p. 695-702.
273. Emami, K.H., W.W. Navarre, and S.T. Smale, *Core promoter specificities of the Sp1 and VP16 transcriptional activation domains*. Mol Cell Biol, 1995. **15**(11): p. 5906-16.
274. Ohtsuki, S., M. Levine, and H.N. Cai, *Different core promoters possess distinct regulatory activities in the Drosophila embryo*. Genes Dev, 1998. **12**(4): p. 547-56.
275. Emami, K.H., A. Jain, and S.T. Smale, *Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization*. Genes Dev, 1997. **11**(22): p. 3007-19.
276. Butler, J.E. and J.T. Kadonaga, *Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs*. Genes Dev, 2001. **15**(19): p. 2515-9.
277. Yean, D. and J. Gralla, *Transcription reinitiation rate: a special role for the TATA box*. Mol Cell Biol, 1997. **17**(7): p. 3809-16.
278. Wefald, F.C., B.H. Devlin, and R.S. Williams, *Functional heterogeneity of mammalian TATA-box sequences revealed by interaction with a cell-specific enhancer*. Nature, 1990. **344**(6263): p. 260-2.
279. Zabidi, M.A., et al., *Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation*. Nature, 2015. **518**(7540): p. 556-9.
280. Bergman, D.T., et al., *Compatibility rules of human enhancer and promoter sequences*. Nature, 2022.
281. Kragestein, B.K., et al., *Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis*. Nat Genet, 2018. **50**(10): p. 1463-1473.
282. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells*. Cell, 2006. **125**(2): p. 315-26.
283. Calo, E. and J. Wysocka, *Modification of enhancer chromatin: what, how, and why?* Mol Cell, 2013. **49**(5): p. 825-37.
284. Hnisz, D., et al., *Super-enhancers in the control of cell identity and disease*. Cell, 2013. **155**(4): p. 934-47.
285. Kvon, E.Z., et al., *Enhancer redundancy in development and disease*. Nat Rev Genet, 2021. **22**(5): p. 324-336.
286. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells*. Cell, 2007. **130**(1): p. 77-88.
287. Hezroni, H., et al., *Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species*. Cell Rep, 2015. **11**(7): p. 1110-22.
288. Pauli, A., et al., *Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis*. Genome Res, 2012. **22**(3): p. 577-91.
289. Bergmann, J.H., et al., *Regulation of the ESC transcriptome by nuclear long noncoding RNAs*. Genome Res, 2015. **25**(9): p. 1336-46.
290. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression*. Nature, 2009. **459**(7243): p. 108-12.
291. Xi, H., et al., *Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome*. PLoS Genet, 2007. **3**(8): p. e136.
292. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals*. Nature, 2009. **458**(7235): p. 223-7.

293. Gil, N. and I. Ulitsky, *Regulation of gene expression by cis-acting long non-coding RNAs*. *Nat Rev Genet*, 2020. **21**(2): p. 102-117.
294. Quinn, J.J. and H.Y. Chang, *Unique features of long non-coding RNA biogenesis and function*. *Nat Rev Genet*, 2016. **17**(1): p. 47-62.
295. Fatica, A. and I. Bozzoni, *Long non-coding RNAs: new players in cell differentiation and development*. *Nat Rev Genet*, 2014. **15**(1): p. 7-21.
296. Liu, L. and F. Fang, *Long Noncoding RNA Mediated Regulation in Human Embryogenesis, Pluripotency, and Reproduction*. *Stem Cells Int*, 2022. **2022**: p. 8051717.
297. Bouckenheimer, J., et al., *Long non-coding RNAs in human early embryonic development and their potential in ART*. *Hum Reprod Update*, 2016. **23**(1): p. 19-40.
298. Qiu, J., et al., *RNA editing regulates lncRNA splicing in human early embryo development*. *PLoS Comput Biol*, 2021. **17**(12): p. e1009630.
299. Schmitz, S.U., P. Grote, and B.G. Herrmann, *Mechanisms of long noncoding RNA function in development and disease*. *Cell Mol Life Sci*, 2016. **73**(13): p. 2491-509.
300. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs*. *Annu Rev Biochem*, 2012. **81**: p. 145-66.
301. Haswell, J.R., et al., *Genome-wide CRISPR interference screen identifies long non-coding RNA loci required for differentiation and pluripotency*. *PLoS One*, 2021. **16**(11): p. e0252848.
302. Jiang, W., et al., *The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression*. *Cell Rep*, 2015. **11**(1): p. 137-48.
303. Yang, J., et al., *GATA6-AS1 Regulates GATA6 Expression to Modulate Human Endoderm Differentiation*. *Stem Cell Reports*, 2020. **15**(3): p. 694-705.
304. Chen, Y.F., et al., *Control of matrix stiffness promotes endodermal lineage specification by regulating SMAD2/3 via lncRNA LINC00458*. *Sci Adv*, 2020. **6**(6): p. eaay0264.
305. Chang, K., S.J. Elledge, and G.J. Hannon, *Lessons from Nature: microRNA-based shRNA libraries*. *Nat Methods*, 2006. **3**(9): p. 707-14.
306. Adamson, B., et al., *A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response*. *Nat Cell Biol*, 2012. **14**(3): p. 318-28.
307. Jackson, A.L., et al., *Expression profiling reveals off-target gene regulation by RNAi*. *Nat Biotechnol*, 2003. **21**(6): p. 635-7.
308. Sigoillot, F.D., et al., *A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens*. *Nat Methods*, 2012. **9**(4): p. 363-6.
309. Engreitz, J.M., et al., *The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome*. *Science*, 2013. **341**(6147): p. 1237973.
310. Shibayama, Y., S. Fanucchi, and M.M. Mhlanga, *Visualization of Enhancer-Derived Noncoding RNA*. *Methods Mol Biol*, 2017. **1468**: p. 19-32.
311. Cabili, M.N., et al., *Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution*. *Genome Biol*, 2015. **16**: p. 20.
312. Lewandowski, J.P., et al., *The Firre locus produces a trans-acting RNA molecule that functions in hematopoiesis*. *Nat Commun*, 2019. **10**(1): p. 5137.
313. Clemson, C.M., et al., *An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles*. *Mol Cell*, 2009. **33**(6): p. 717-26.

314. Andergassen, D. and J.L. Rinn, *From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo*. *Nat Rev Genet*, 2022. **23**(4): p. 229-243.
315. Davenport, C., et al., *Anterior-Posterior Patterning of Definitive Endoderm Generated from Human Embryonic Stem Cells Depends on the Differential Signaling of Retinoic Acid, Wnt-, and BMP-Signaling*. *Stem Cells*, 2016. **34**(11): p. 2635-2647.
316. Gaertner, B., et al., *A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation*. *Elife*, 2020. **9**.
317. Nowotschin, S., et al., *The emergent landscape of the mouse gut endoderm at single-cell resolution*. *Nature*, 2019. **569**(7756): p. 361-367.
318. Linneberg-Agerholm, M., et al., *Naive human pluripotent stem cells respond to Wnt, Nodal and LIF signalling to produce expandable naive extra-embryonic endoderm*. *Development*, 2019. **146**(24).
319. Viotti, M., S. Nowotschin, and A.K. Hadjantonakis, *SOX17 links gut endoderm morphogenesis and germ layer segregation*. *Nat Cell Biol*, 2014. **16**(12): p. 1146-56.
320. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system*. *Nat Protoc*, 2013. **8**(11): p. 2281-2308.
321. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems*. *Science*, 2013. **339**(6121): p. 819-23.
322. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications*. *Nature*, 2016. **538**(7624): p. 265-269.
323. Weintraub, A.S., et al., *YY1 Is a Structural Regulator of Enhancer-Promoter Loops*. *Cell*, 2017. **171**(7): p. 1573-1588.e28.
324. Genga, R.M.J., et al., *Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development*. *Cell Rep*, 2019. **27**(3): p. 708-718 e10.
325. Yin, Y., et al., *Opposing Roles for the lncRNA Haunt and Its Genomic Locus in Regulating HOXA Gene Activation during Embryonic Stem Cell Differentiation*. *Cell Stem Cell*, 2015. **16**(5): p. 504-16.
326. Stewart, S.A., et al., *Lentivirus-delivered stable gene silencing by RNAi in primary cells*. *RNA*, 2003. **9**(4): p. 493-501.
327. Rueden, C.T., et al., *ImageJ2: ImageJ for the next generation of scientific image data*. *BMC Bioinformatics*, 2017. **18**(1): p. 529.
328. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis*. *Nat Methods*, 2012. **9**(7): p. 676-82.
329. Servant, N., et al., *HiC-Pro: an optimized and flexible pipeline for Hi-C data processing*. *Genome Biol*, 2015. **16**: p. 259.
330. Schmitt, A.D., et al., *A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome*. *Cell Rep*, 2016. **17**(8): p. 2042-2059.
331. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. *Bioinformatics*, 2010. **26**(6): p. 841-2.
332. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif*. *Bioinformatics*, 2011. **27**(7): p. 1017-8.
333. Mathelier, A., et al., *JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles*. *Nucleic Acids Res*, 2016. **44**(D1): p. D110-5.

334. Liu, T., et al., *Cistrome: an integrative platform for transcriptional regulation studies*. *Genome Biol*, 2011. **12**(8): p. R83.
335. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes*. *Nature*, 2015. **518**(7539): p. 317-30.
336. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS)*. *Genome Biol*, 2008. **9**(9): p. R137.
337. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. *J EMBnet.journal*. 2011. **17**(1): p. 10.
338. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. *Nat Methods*, 2012. **9**(4): p. 357-9.
339. Kent, W.J., et al., *The human genome browser at UCSC*. *Genome Res*, 2002. **12**(6): p. 996-1006.
340. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes*. *Nucleic Acids Res*, 2019. **47**(D1): p. D766-D773.
341. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv e-prints, 2013: p. arXiv:1303.3997.
342. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*. *Genome Res*, 2010. **20**(9): p. 1297-303.
343. Athar, A., et al., *ArrayExpress update - from bulk to single-cell expression data*. *Nucleic Acids Res*, 2019. **47**(D1): p. D711-D715.
344. Frankish, A., et al., *Gencode 2021*. *Nucleic Acids Res*, 2021. **49**(D1): p. D916-D923.
345. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression*. *Nat Methods*, 2017. **14**(4): p. 417-419.
346. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis*. *Genome Biol*, 2018. **19**(1): p. 15.
347. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. *Bioinformatics*, 2013. **29**(1): p. 15-21.
348. Robinson, J.T., et al., *Integrative genomics viewer*. *Nat Biotechnol*, 2011. **29**(1): p. 24-6.
349. Giesselmann, P., et al., *Nanopype: a modular and scalable nanopore data processing pipeline*. *Bioinformatics*, 2019. **35**(22): p. 4770-4772.
350. Li, H., *Minimap2: pairwise alignment for nucleotide sequences*. *Bioinformatics*, 2018. **34**(18): p. 3094-3100.
351. Martin, M., *Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads*. *EMBnet.journal*, 2011. **17**(1): p. 10-12.
352. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads*. *Nat Biotechnol*, 2015. **33**(3): p. 290-5.
353. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. *Genome Biol*, 2014. **15**(12): p. 550.
354. Durand, N.C., et al., *Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom*. *Cell Syst*, 2016. **3**(1): p. 99-101.
355. Zhou, X., et al., *Exploring long-range genome interactions using the WashU Epigenome Browser*. *Nat Methods*, 2013. **10**(5): p. 375-6.
356. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

357. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
358. Kuhn, R.M., D. Haussler, and W.J. Kent, *The UCSC genome browser and associated tools*. Brief Bioinform, 2013. **14**(2): p. 144-61.

## 7 MATERIALS AND METHODS

### 7.1 Experimental Materials and Approaches

#### 7.1.1 Molecular cloning

##### Molecular cloning of boundary knock-out constructs

For CRISPR/Cas9 mediated targeting of either *SOX17* or *NANOG* boundary knock-out constructs we utilized pSpCas9(BB)-2A-GFP (PX458), which was a gift from Feng Zhang (Addgene plasmid # 48138 ; <http://n2t.net/addgene:48138> ; RRID:Addgene\_48138)<sup>[320]</sup>. Prior to small guide RNA (sgRNA) cloning, pX458 was initially modified and further re-named into 2X\_pX458. 2X\_pX458 harbors an additional independent U6-promoter followed by a small guide RNA (sgRNA) scaffold expression cassette, which allows the insertion of an additional sgRNA by Sapl restriction enzyme cloning. To generate 2X\_pX458, pX458 and the synthesized Sapl sgRNA expression cassette (IDT, [find oligonucleotide sequence below](#)) were digested with KpnI (New England Biolabs, R3142S). Next, the Sapl sgRNA expression cassette was ligated into the KpnI linearized pX458 in a 3:1 molarity ratio using T4 DNA-ligase (New England Biolabs, M0202S) according to the manufacturer's instructions followed by transformation and Sanger sequencing to verify successful cloning.

sgRNA-cloning was performed with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, E2621S) according to manufacturer's instructions using BbsI-linearization of 2X\_pX458 for the first sgRNA and Sapl linearization of 2X\_pX458 for the second sgRNA as backbone, combined with single stranded oligonucleotides containing the sgRNA sequences as inserts (1:3 molar ratio) ([find sgRNA sequences below](#)). Bacterial transformation and Sanger sequencing was performed to verify successful cloning. Empty 2X\_pX458 was deposited on addgene.org under ID #172221. The 2X\_pX458 derived *SOX17* and *NANOG* boundary knock-out constructs were deposited on addgene.org under ID #172225 and ID #172224 respectively.

Oligonucleotide Name	Purpose	5'-3' Sequence
Sapl_sgRNA_case tte_gBlock_+KpnI- sites	Extension cloning of PX458	tgcagacaaatggctctagaggtacggtaccAATATGCATT TTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTGGAATTAAT TTGACTGTAAACACAAAGATATTAGTACAAAAT ACGTGACGTAGAAAGTAATAATTTCTTGGGTA GTTTGCAGTTTTTAAAATTATGTTTTAAAATGGA CTATCATATGCTTACCGTAACTTGAAAGTATTT



CGATTTCTTGGCTTTATATATCTTGTGGAAAG  
 GACGAAACACCGGAAGAGCGAGCTCTTCTGT  
 TTTAGAGCTAGAAATAGCAAGTTAAAATAAGG  
 CTAGTCCGTTATCAACTTGAAAAAGTGGCACC  
 GAGTCGGTGCTTTTTTGTTCCTGCAGGAGA  
 TTTAgcgcgtgccaattctgcagacaaatggctctagaggta  
 cccgttacataacttacggtaaatggA

	CRISPR/Cas9	
AL_SOX17_CTCF	mediated	TGTGGAAAGGACGAAACACCCACATCCAGTCT
_left_gRNA_1	targeting	GCCAACATAGTTTTAGAGCTAGAAATAGC
	/SOX17	
	Boundary 2	
	CRISPR/Cas9	
AL_SOX17_CTCF	mediated	TGTGGAAAGGACGAAACACCGGGCTGCACCA
_left_gRNA_2	targeting	AATCGCCACGTTTTAGAGCTAGAAATAGC
	/SOX17	
	Boundary 2	

### Molecular cloning of SOX17 and eSOX17.2 knock-out constructs

For CRISPR/Cas9 mediated targeting of either *SOX17* or *eSOX17.2* we utilized pSpCas9(BB)-2A-GFP [320] (PX458), which was a gift from Feng Zhang (Addgene plasmid # 48138 ; <http://n2t.net/addgene:48138> ; RRID:Addgene\_48138) [320]. Prior to small guide RNA (sgRNA) cloning, PX458 was initially modified and further re-named into P2X458. P2X458 harbors an additional independent U6-promoter followed by a small guide RNA (sgRNA) scaffold expression cassette, which allows the insertion of an additional sgRNA by Sapl restriction enzyme cloning. To generate P2X458, PX458 and the synthesized Sapl sgRNA expression cassette (IDT, find sequence under 7.1.2) were digested with KpnI (New England Biolabs, R3142S). Next, the Sapl sgRNA expression cassette was ligated into the KpnI linearized PX458 in a 3:1 molarity ratio using T4 DNA-ligase (New England Biolabs, M0202S) according to the manufacturer's instructions followed by transformation and Sanger sequencing to verify successful cloning. sgRNA-cloning was performed with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, E2621S) according to manufacturer's instructions using BbsI-linearization of P2X458 for the first sgRNA and Sapl linearization of P2X458 for the second sgRNA as backbone, combined with single stranded oligonucleotides containing the sgRNA sequences as inserts (1:3 molar ratio) ([find sgRNA sequences below](#)). Bacterial transformation and Sanger sequencing was performed to verify successful cloning.

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
SOX17_cKO_sgRNA_1	CRISPR/Cas9 mediated targeting /SOX17 gene body	AGCTCCGGCTAGTTTTCCCG
SOX17_cKO_sgRNA_2	CRISPR/Cas9 mediated targeting /SOX17 gene body	ACGTGCCACAGCGATTAGTA
eSOX17.2_KO_sgRNA_left	CRISPR/Cas9 mediated targeting /eSOX17.2	GATTAGGTGGCCCCTAACAC
eSOX17.2_KO_sgRNA_right	CRISPR/Cas9 mediated targeting /eSOX17.2	CAAATCCATGCTAGGCTCC

### Molecular cloning of Luciferase reporter constructs

pGL4.27[luc2P/minP/Hygro] (Promega, E8451) containing a minimal CMV-promoter for enhancer-assays or pGL4.15[luc2P/Hygro] (Promega, E6701) w/o any promoter for promoter-assays were first digested using EcoRV (New England Biolabs, R3195S). Next, full *eSOX17*, *eSOX17.1* or *eSOX17.2* for enhancer-assays and *pSOX17* or *pLNCSOX17* genomic regions were PCR amplified with primers containing homology overhangs to the plasmid ([find primer sequences below](#)). PCR products were purified and cloned into the linearized plasmid utilizing the NEBuilder HiFi DNA Assembly Master Mix (1:3 molar ratio) according to the manufacturer's instructions. Bacterial transformation followed by Sanger sequencing verified the successful cloning.

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
eSOX17.1+2_fwd	Overhang PCR - Luciferase cloning /eSOX17.1+2	ACCTGAGCTCGCTAGCCTC GAGGATATTTAAAGCTTCGT AGCATCAGGTGTG
eSOX17.1+2_rev	Overhang PCR - Luciferase cloning /eSOX17.1+2	TTGGCCGCCGAGGCCAGAT CTTGATTCCAGGAACACAAA ATAAGCAAGTGCC

eSOX17.1_fwd	Overhang PCR - Luciferase cloning /eSOX17.1	ACCTGAGCTCGCTAGCCTC GAGGATATTTAAAGCTTCGT AGCATCAGGTGTG
eSOX17.1_rev	Overhang PCR - Luciferase cloning /eSOX17.1	TTGGCCGCCGAGGCCAGAT CTTGATGTTAGGGGCCACC TAATCAATGCC
eSOX17.2_fwd	Overhang PCR - Luciferase cloning /eSOX17.2	ACCTGAGCTCGCTAGCCTC GAGGATACTGGGGGTCAAT CTGTCCAG
eSOX17.2_rev	Overhang PCR - Luciferase cloning /eSOX17.2	TTGGCCGCCGAGGCCAGAT CTTGATTCCAGGAACACAAA ATAAGCAAGTGCC
pSOX17_fwd	Overhang PCR - Luciferase cloning /pSOX17	ACCTGAGCTCGCTAGCCTC GAGGATATAACAAAATGCAC ACCTTGCCCTAAC
pSOX17_rev	Overhang PCR - Luciferase cloning /pSOX17	TTGGCCGCCGAGGCCAGAT CTTGATGCCGGCCTAGTGA CACTG
pLNC SOX17_fwd	Overhang PCR - Luciferase cloning /pLNC SOX17	ACCTGAGCTCGCTAGCCTC GAGGATCCCAAATCCCCCA AACATTACAAC
pLNC SOX17_rev	Overhang PCR - Luciferase cloning /pLNC SOX17	TTGGCCGCCGAGGCCAGAT CTTGATAAATGAAGGGAAAA TGTGGAAAACCTGG

### Molecular cloning of lentiviral sgRNA constructs

pU6-sgRNA EF1Alpha-puro-T2A-BFP [229] was digested with BstXI (New England Biolabs, R0113S) and BlnI (New England Biolabs, R0585S) and the linearized plasmid was gel extracted with the QIAquick Gel Extraction Kit (Qiagen, 28704). Subsequently sgRNAs (sgLNC SOX17 or sgCtrl) ([find sgRNA sequences below](#)) were cloned in the linearized backbone using NEBuilder HiFi DNA Assembly Master Mix (1:3 molar ratio) according to the manufacturer's instructions. Bacterial transformation and sanger sequencing confirmed the

successful cloning. pU6-sgRNA EF1Alpha-puro-T2A-BFP was a gift from Jonathan Weissman (Addgene plasmid # 60955 ; <http://n2t.net/addgene:60955> ; RRID:Addgene\_60955).

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
LNC $SOX17$ _sgRNA	dCas9-KRAB-MeCP2 repression /-355 bp TSS <i>LNC<math>SOX17</math></i>	ataagtatcccttggagaaccacctgttg AGTGGTGTGGATTTTCGGCA GGTTTAAGAGCTATGCTGG AAACAGCaTAGCAAGTTTAA AT
Ctrl_sgRNA	dCas9-KRAB-MeCP2 repression /unrelated target (Gilbert et al. 2014)	ataagtatcccttggagaaccacctgttg GCGCCAAACGTGCCCTGAC GGTTTAAGAGCTATGCTG GAAACAGCaTAGCAAGTTTA AAT

### Molecular cloning of $SOX17$ reporter knock-in constructs

pUC19 plasmid was digested with *Sma*I (New England Biolabs, R0141S) and the linearized plasmid was gel extracted with the QIAquick Gel Extraction Kit (Quiagen, 28704). Next,  $SOX17$  homology arm genomic regions were PCR amplified with primers containing homology overhangs ([find primer sequences listed below](#)) to the plasmid and to a T2A-H2B-mCitrine-loxP-hPGK-BSD-loxP selection cassette.

The left homology arm overlapped with the end of the  $SOX17$  coding sequence, and the T2A-H2B-mCitrine cassette which was cloned in frame with the last  $SOX17$  aminoacid. PCR products and selection cassette were purified and cloned into the linearized plasmid utilizing the NEBuilder HiFi DNA Assembly Master Mix according to the manufacturer's instructions. Bacterial transformation followed by Sanger sequencing verified the successful cloning.

sgRNA targeting the genomic region of integration was cloned in *Bbs*I linearized pX335-U6-Chimeric\_BB-CBh-hSpCas9n [321] (D10A) plasmid (Addgene #42335) using NEBuilder HiFi DNA Assembly Master Mix (1:3 molar ratio) ([find sgRNA sequence listed below](#)) according to the manufacturer's instructions. pX335-U6-Chimeric\_BB-CBh-hSpCas9n(D10A) was a gift from Feng Zhang (Addgene plasmid # 42335 ; <http://n2t.net/addgene:42335> ; RRID:Addgene\_42335) Bacterial transformation and sanger sequencing confirmed the successful cloning.

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
SOX17_HA_left_fwd	Cloning of SOX17 HA into pUC19 plasmid for SOX17 reporter /SOX17 C-terminal	tttgctggccttttgctcacatgtGGGC CTGGAGCGGGAGCGCA
SOX17_HA_left_rev	Cloning of SOX17 HA into pUC19 plasmid for SOX17 reporter /SOX17 C-terminal	CTCTGCCCTCTCCACTGCC GAATTCCACGTCAGGATAGT TGCAGT
SOX17_HA_right_fwd	Cloning of SOX17 HA into pUC19 plasmid for SOX17 reporter /SOX17 C-terminal	TTATACGAAGTTATGGCGCG CCAGCCAGGTCCCTGATCC GCCCA
SOX17_HA_right_rev	Cloning of SOX17 HA into pUC19 plasmid for SOX17 reporter /SOX17 C-terminal	acctctgacacatgcagctcccggaAA CCATTCATGGATTCTCCC
SOX17_gRNA	CRISPR/Cas9 mediated targeting /SOX17 C-terminal	TGTGGAAAGGACGAAACAC CGCAGTAATATACCGCGGA GCGTTTTAGAGCTAGAAATA GC

### Molecular cloning of *LNC*SOX17-promoter-KI constructs

pUC19 plasmid was digested with *Sma*I (New England Biolabs, R0141S) and the linearized plasmid was gel extracted with the QIAquick Gel Extraction Kit (Quiagen, 28704). Next, *LNC*SOX17 homology arm genomic regions were PCR amplified with primers ([find primer sequences below](#)) containing homology overhangs to the plasmid and to a mRuby-3xFLAG-NLS-3xSV40-poly(A)-loxP-mPGK-PuroR-loxP selection cassette.

The left homology arm overlapped with the *LNC*SOX17 promoter including 30 bp of *LNC*SOX17 Exon 1, and a mRuby-3xFLAG-NLS-3xSV40-poly(A) cassette which was cloned +30 bp after *LNC*SOX17-TSS into Exon 1. The right homology arm overlapped with *LNC*SOX17 Exon 1 -30 bp TSS, and a loxP-mPGK-PuroR-loxP cassette which was cloned following the mRuby-3xFLAG-NLS-3xSV40-poly(A) cassette. Both, the mRuby-3xFLAG-NLS-3xSV40-poly(A) and the loxP-mPGK-PuroR-loxP cassette also shared homology. All PCR products were purified and cloned into the linearized plasmid utilizing the NEBuilder HiFi DNA Assembly Master Mix according to the manufacturer's instructions. Bacterial transformation followed by Sanger sequencing verified the successful cloning.

For CRISPR/Cas9 mediated targeting of the *LNC SOX17* promoter we utilized pSpCas9(BB)-2A-GFP [320] (PX458), [ENREF\\_1](#) which was a gift from Feng Zhang (Addgene plasmid # 62988 ; <http://n2t.net/addgene:62988> ; RRID:Addgene\_62988) [320]. sgRNA-cloning was performed with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, E2621S) according to manufacturer's instructions using BbsI-linearization of PX458, combined with single stranded oligonucleotides containing the sgRNA sequences as inserts (1:3 molar ratio) ([find sgRNA sequence below](#)). Bacterial transformation and Sanger sequencing was performed to verify successful cloning.

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
LNC SOX17_HA_left_F	Cloning homology arm (left) PCR - <i>LNC SOX17</i> - promoter-KI cassette /-30 bp <i>LNC SOX17</i> TSS	tgcaggctcgactctagaggatccc cTTGAGCAGCTGGCCTG GGGTCA
LNC SOX17_HA_left_R	Cloning homology arm (left) PCR - <i>LNC SOX17</i> - promoter-KI cassette /-30 bp <i>LNC SOX17</i> TSS	TCCTCTCACACCTGGG CTG
LNC SOX17_Puro_cassette_F	Cloning PuroR selection cassette PCR - <i>LNC SOX17</i> -promoter-KI cassette /-30 bp <i>LNC SOX17</i> TSS	GGTGGCAGCCCAGGTG GTGAGAGGAGCTGGCG CGCCATAACTTCG
LNC SOX17_Puro_cassette_R	Cloning PuroR selection cassette PCR - <i>LNC SOX17</i> -promoter-KI cassette /-30 bp <i>LNC SOX17</i> TSS	ATAGTACTTAAATAACTT CG
mRuby_F	Cloning mRuby-3xpoly(A) cassette PCR - <i>LNC SOX17</i> -promoter-KI cassette /-30 bp <i>LNC SOX17</i> TSS	CTATACGAAGTTATTTA AGTACTATTGGAGACTA ACTTGTTTATT

mRuby_R	Cloning mRuby-3xpoly(A) cassette PCR - LNC SOX17-promoter-KI cassette /-30 bp LNC SOX17 TSS	gccaccatggtgtctaaggg
LNC SOX17_HA_right_F	Cloning homology arm (right) PCR - LNC SOX17- promoter-KI cassette /-30 bp LNC SOX17 TSS	cttcgcccttagacaccatggtggc ACCATCTTCAGGGGGA CAGA
LNC SOX17_HA_right_R	Cloning homology arm (right) PCR - LNC SOX17- promoter-KI cassette /-30 bp LNC SOX17 TSS	gccagtgaattcgagctcgggtacc cCAGCCTGGGCAACATG GTGA
LNC SOX17_STOP_KI_sgRNA	CRISPR/Cas9 mediated targeting/ LNC SOX17- promoter -30bp TSS /-30 bp LNC SOX17 TSS	TGTGGAAAGGACGAAA CACCGTGAGAGGAACC ATCTTCAGGGTTTTAGA GCTAGAAATAGC

### 7.1.2 Sequencing and library preparations

#### Hi-C sequencing

Hi-C libraries were prepared following the protocol described in Rao et. al. 2014[109]. Briefly, one million cells were crosslinked with final 1% formaldehyde (Thermo Fischer Scientific, 28908) for 10 minutes at room temperature and then quenched with final 0.2 M glycine (Sigma-Aldrich, 50046) solution. Cells were lysed and nuclei permeabilized with 0.5% sodium dodecyl sulphate (Thermo Fisher Scientific, AM9820) for 10 minutes at 62°C. Chromatin was digested with 100 U of MboI restriction enzyme (New England Biolabs, R0147L). Ends of the restriction fragments were blunted and labeled with a biotinylated nucleotide and then ligated. Nuclei were pelleted, proteins were digested with proteinase K and crosslinks were reversed by heating at 68°C overnight. DNA was sheared in a Covaris focused ultrasonicator to average fragment length of 400 bp. Size-selected DNA was enriched for biotinylated ligation products through binding to T1 streptavidin beads (Thermo Fisher, 65601). Libraries were prepared for Illumina sequencing by performing the end-repair, A-tailing and adapter ligation steps with DNA attached to the beads. Hi-C libraries were amplified directly off the beads and purified for subsequent Illumina sequencing with 100 paired-ends.

### **SureSelect cHi-C probe Design**

The library of SureSelect enrichment probes were designed over the genomic interval (hg19, chr8:54735936-55657612) using the SureDesign online tool of Agilent. 3299 total probes cover the *SOX17* locus and were designed to specifically enrich for regions in proximity of *Nla*III sites. The probes covered 35,25% of the interval. Probe sequences can be requested under the SureSelect DNA design ID 3253271 from Agilent Technologies.

### **Capture Hi-C (cHi-C) sequencing**

cHi-C libraries were either prepared from wild type/homozygous *SOX17*<sup>Δ5<sup>CTCF</sup></sup> iPSC/EN or CRISPRi sgCtrl/sgLNC*SOX17* EN cells. Undifferentiated or day 5 differentiated ZIP13K2[211] cells were grown to a final count of 4-5 million, treated with Accutase (Sigma-Aldrich, A6964), resuspended and washed in DPBS. Cell lysis, *Nla*III (NEB R0125) digestion, ligation and de-crosslinking was performed according to the Franke et al. protocol[322]. Adaptors were added to DNA and amplified according to Agilent instructions for Illumina sequencing. The library was hybridized to the custom-designed sure-select beads and indexed for sequencing of 200x10<sup>6</sup> fragments per sample (100 bp paired-end) following the Agilent instructions. Capture Hi-C experiments were performed as biological duplicates.

### **RNA sequencing – *SOX17* CTCF loop domain perturbations studies**

Triplicates of either undifferentiated or differentiated ZIP13K2 cultures were treated with Accutase (Sigma-Aldrich, A6964) and differentiated cultures were further quenched with FACS-buffer containing 5 mM EDTA (ThermoFischer Scientific, 15575020) 10% FBS (ThermoFischer Scientific, 26140079) in DPBS (Thermo Fischer Scientific, 14190250) to obtain single cells. In order to enrich for CXCR4<sup>-</sup> or CXCR4<sup>+</sup> cell fractions of differentiated cultures, cells were stained for anti-Human CXCR4 (CD184) PE (as described under 21. FACS) and compared to Isotype and unstained control sorted for either CXCR4<sup>-</sup> or CXCR4<sup>+</sup> sub-populations on the Aria II (Beckton Dickinson). RNA isolation including on-column DNase digest of enriched cell populations was performed using the RNeasy Mini Kit (Qiagen, 74104) according to the manufacturer's instructions. The KAPA Stranded mRNA-Seq Kit (Kapa Biosystems, #KK8401) was utilized for RNA library preparation, using 500 ng total RNA and performing poly-(A) enrichment followed by first strand cDNA-synthesis (11 cycles). Subsequently, RNA sequencing libraries were prepared by the use of dual index primers according to the manufacturer's instructions. Illumina adapter ligated sequencing libraries were sequenced for 50 million 75 bp long read pairs per sample on the HiSeq4000 (Illumina).

### **RNA sequencing – *LNC**SOX17* repression studies**

ZIP13K2 hiPSCs and their derived EN cultures were treated with Accutase for 15 min at 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cells were then collected, washed with ice cold



DPBS and centrifuged at 4°C, 300 x g for 5 min. Subsequently, 350 µl of RLT Plus buffer containing 1% β-mercaptoethanol (Thermo) was added to the cell pellets for cell lysis. After dissociation by trituration and vortexing, RNA was extracted using RNeasy Plus Micro Kit (Qiagen) and RNA concentration and quality was measured using the Agilent RNA 6000 Pico Kit (Agilent Technologies, 5067-1513) on an Agilent 2100 Bioanalyzer. All samples analyzed had a RINe value higher than 8,0 and were subsequently used for library preparation. mRNA libraries were prepared using KAPA Stranded RNA-Seq Kit (KapaBiosystem) according to the manufacturer's instructions. 500 ng of total RNA was used for each sample to enter the library preparation protocol. For adapter ligation dual indexes were used (NEXTFLEX® Unique Dual Index Barcodes NOVA-514150) at a working concentration of 71nM (5 µl of 1 uM stock in each 70 µl ligation reaction). Quality and concentration of the obtained libraries were measured using Agilent High Sensitivity D5000 ScreenTape (Agilent-Technologies, 5067- 5592) on an Agilent 4150 TapeStation. All libraries were sequenced using 100 bp paired-end sequencing (200 cycles kit) on a NovaSeq platform at a minimum of 25 million fragments /sample.

### **Extraction of poly(A)-RNA for Nanopore long-read sequencing**

Isolation of poly(A)-enriched mRNA was performed using the Dynabeads mRNA DIRECT purification kit (Thermo Fisher Scientific, 61011) according to the manufacturer's instruction with minor modifications. ZIP13K2-derived EN cells were washed once with DPBS and dissociated with Accutase for 15 min at 37°C, 5% CO<sub>2</sub>. Enzymatic reaction was quenched by adding mTeSR1 and cells were counted using the Countess II automated cell-counter. A total of 4 x 10<sup>6</sup> viable cells were centrifuged for 5 min at 4°C, 300 x g. The supernatant was discarded, and cells were washed with 1 ml of ice-cold DPBS and centrifuged as described above. The supernatant was completely removed, and the cell pellet was carefully resuspended in 1,25 ml Lysis/Binding buffer. In order to reduce viscosity resulting from released genomic DNA, the samples were passed through a 21-gauge needle (Becton Dickinson, 304432) for five times and subsequently added to the pre-washed Oligo(dT)<sub>25</sub> beads. Hybridization of the beads/mRNA complex was carried out for 10 min on a Mini Rotator (Grant-bio) and vials were placed on a DynaMag2 magnet (Thermo Fisher Scientific, 12321D) until the beads were fully immobilized. The DNA containing supernatant was removed and the beads were resuspended twice with 2 ml of Buffer A following a second wash step with two times 1 ml of Buffer B. Purified RNA was eluted with 10 µl of pre-heated Elution Buffer (10 mM Tris-HCl pH 7,5) for 5 min at 80°C and quantified with a Qubit Fluorometer (Thermo Fisher Scientific) using the RNA HS Assay Kit (Thermo Fisher Scientific, Q32852). Eluted RNA samples were immediately used for preparation of Nanopore sequencing libraries or kept at -80°C.

### Preparation of Nanopore sequencing libraries

Preparation of RNA sequencing libraries was performed following the manufacturer's instructions (ONT, SQK-PCS109) with minor modifications. Briefly, 50 ng of freshly prepared poly(A)-enriched mRNA was subjected to reverse transcription and strand-switching reaction. A total of four PCR reactions, each containing 5 µl of reverse transcribed cDNA, was used for the attachment of rapid primers (cPRM). Sufficient amplification of long cDNA molecules was enabled by setting the PCR extension time to 19 min and a total of 12 x cycles were used for amplification. Samples were treated with 1 µl of Exonuclease I (New England Biolabs, M0293S) and subsequently pooled for SPRI bead cleanup. Wash steps were performed using 80% ethanol solution and beads were eluted in 60 µl of 50°C pre-heated nuclease-free water. Samples were then incubated for additional 20 min at 50°C. Eluted DNA was combined with 5 µl adapter mix (AMX), 25 µl ligation buffer (LNB) from ONTs ligation sequencing kit (ONT, SQK-LSK109) and 10 µl of NEBNext Quick T4 DNA Ligase (New England Biolabs, E6056S). Ligation mix was incubated at RT for 30 min. Removal of short DNA fragments was achieved by adding 40 µl of Agencourt AMPure XP beads (Beckmann Coulter, A63881) combined with 2 wash steps with 250 µl of long fragment buffer (LFB) included in ONTs ligation sequencing kit. The final library was eluted with 13 µl elution buffer (EB) for 20 min at 48°C and DNA concentration was quantified using the Qubit dsDNA BR assay kit (Thermo Fisher Scientific, Q32850). A total of 400 ng was carefully mixed with 37,5 µl sequencing buffer (SQB), 25,5 µl of loading beads (LB) and loaded onto a primed MinION flow cell (ONT, R9.4.1 FLO-MIN106).

### 4C sequencing

Triplicates of either undifferentiated ZIP13K2 or ZIP13K2-derived EN cultures were collected as described previously. ZIP13K2-derived EN cultures were further quenched with MACS-buffer (Final DPBS, 2 mM EDTA (ThermoFisher Scientific), 0,5% BSA (Sigma-Aldrich)) to obtain a single cell suspension. CXCR4<sup>+</sup> cell populations, were enriched using MicroBead Kit (Miltenyi Biotec) following the manufacturer's instructions. Pre- and post-MACS enriched cell fractions of differentiated cultures were measured for CXCR4-APC signal on the FACS Aria II (Beckton Dickinson) to confirm the cell population purity. Circularized Chromosome Conformation Capture (4C) library preparation of undifferentiated, or differentiated CXCR4<sup>+</sup> enriched cell populations was performed according to the Weintraub A.S. et al. protocol [323]. Briefly, NlaIII (New England Biolabs, R0125) was used as the primary cutter and DpnII (New England Biolabs, R0543) as a secondary cutter. Touchdown PCR on 4C libraries was performed using specific primer-pairs ([find primer sequences below](#)) for the respective view-point. Illumina sequencing libraries were then prepared and sequenced using 150 paired-end sequencing (300 cycles kit) on a HiSeq4000 platform at a minimum of 10M fragments/ sample.

Oligonucleotide Name	Purpose /Region	5'-3' Sequence
ND23658576_1f	4C sequencing /eSOX17	TAAGACAAAGTATCTCCATG
ND23658576_2r	4C sequencing /eSOX17	CACAACCTCCTATCCAAAGA

### SOX17 Chromatin Immunoprecipitation (ChIP) sequencing

ZIP13K2-derived EN cells ( $5 \times 10^6$  /IP) were harvested and cross-linked in 1% formaldehyde (Thermo Fisher Scientific, 28908) in DPBS for 10 min at RT, followed by quenching with final 125 mM Glycine (Sigma-Aldrich, 50046) for 5 min at RT. Cross-linked cells were then centrifuged at  $500 \times g$  at  $4^\circ\text{C}$  and washed twice with ice cold DPBS. Cell lysis was performed by resuspending the pellet in 500  $\mu\text{l}$  Cell Lysis Buffer (Final 10mM Tris-HCl, pH 8,0 (Sigma Aldrich, T2694); 85mM KCl (Sigma Aldrich, P9541); 0,5% NP40 (Sigma Aldrich, 56741); 1 x cComplete, EDTA-free Protease Inhibitor Cocktail (Sigma Aldrich, 11873580001)) followed by 10 min incubation on ice. After the incubation, lysed cells were centrifuged at  $2500 \times g$  for 5 min at  $4^\circ\text{C}$ . Supernatant was carefully removed and the extracted nuclei were then resuspended in 230  $\mu\text{l}$  Nuclei Lysis Buffer (Final 10mM Tris-HCl, pH 7,5 (Sigma Aldrich, T2319)); 1% NP40; 0,5% sodium deoxycholate (Sigma Aldrich, D6750); 0,1% SDS (Thermo Fisher Scientific, AM9820); 1 x cComplete, EDTA-free Protease Inhibitor Cocktail). Following 10 min incubation on ice, each 260  $\mu\text{l}$  sample was split into two microTUBEs (Covaris, 520045) and chromatin was sonicated using a Covaris E220 Evolution with the following settings: Temperature  $\rightarrow 4^\circ\text{C}$ ; Peak power  $\rightarrow 140$ ; Duty factor  $\rightarrow 5,0$ ; Cycles/Burst  $\rightarrow 200$ ; Duration  $\rightarrow 750$  sec. After sonication, sheared chromatin (ranging from 200-600bp) was transferred in a new 1,5 ml tube and centrifuged at max speed for 10 min at  $4^\circ\text{C}$ . Supernatant was then transferred into a new tube and volume was increased to 1 ml /sample with ChIP Dilution Buffer (Final 16,7mM Tris-HCl, pH 8,0; 1,2mM EDTA (Sigma Aldrich, 03690)); 167mM NaCl (Sigma Aldrich); 1,1% Triton-X (Sigma Aldrich); 0,01% SDS; 1 x Protease Inhibitor). 50 $\mu\text{l}$  (5%) was then transferred into a new tube and frozen at  $-20^\circ\text{C}$  as INPUT. 1 $\mu\text{g}$  of SOX17 antibody / $10^6$  initial cells was added to the 950  $\mu\text{l}$  left, and immunoprecipitation was carried out at  $4^\circ\text{C}$  o/n on a rotator ([find antibody listed below](#)). The next day, 50 $\mu\text{l}$  of Dynabeads Protein G (Thermo Fisher Scientific, 10004D) /IP were washed twice with ice cold ChIP Dilution Buffer and then added to each IPs. IP/bead mixes were incubated for 4 hours at  $4^\circ\text{C}$  on a rotor. Next, bead/chromatin complexes were washed twice with Low Salt Wash Buffer at  $4^\circ\text{C}$  (Final 20 mM Tris-HCl, pH 8,0; 2 mM EDTA; 150 mM NaCl (Sigma-Aldrich, S6546); 1% Triton-X; 0,1% SDS), twice with High Salt Wash Buffer at  $4^\circ\text{C}$  (Final 20 mM Tris-HCl, pH 8,0; 2 mM EDTA; 500 mM NaCl; 1% Triton-X; 0,1% SDS), twice with LiCl Wash Buffer at  $4^\circ\text{C}$  (Final 10 mM Tris-HCl, pH 8,0; 1mM EDTA; 250mM LiCl (Sigma Aldrich, L9650); 1% sodium deoxycholate (Sigma

Aldrich); 1% NP40), twice with TE pH 8,0 (Sigma Aldrich, 8890) at room temperature and finally eluted twice in 50 µl freshly prepared ChIP Elution Buffer (Final 0,5% SDS; 100 mM NaHCO<sub>3</sub> (Sigma Aldrich, S5761)) at 65°C for 15 min (total 100 µl final eluent). Thawed INPUTS and eluted IPs were next reverse cross-linked at 65°C o/n after the addition of 16 µl freshly prepared Reverse Crosslinking Salt Mixture (Final 250 mM Tris-HCl, pH 6,5 (Sigma Aldrich, 20-160); 62,5 mM EDTA; 1,25M NaCl; 5 mg/ml Proteinase K (Thermo Fisher Scientific, AM2548)). The following day phenol:chloroform (Thermo Fisher Scientific, 15593031) extraction followed by precipitation was performed to isolate DNA. IPs and INPUTS were then quantified and NGS libraries were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, #E7645) following the manufacturer's instructions. Library quality and size distribution was verified using a TapeStation D5000 HS kit (Agilent Technologies, 5067-5592). Samples were sequenced with a coverage of 50 M paired end reads (2 x 100 bp) /sample on a NovaSeq (Illumina).

<b>Antibody name</b>	<b>Application /Dilution used</b>	<b>Clone /Company</b>	<b>Catalogue number</b>
Goat anti-Human SOX17 (unconjugated)	1 <sup>st</sup> ChIP /1µg per 10 <sup>6</sup> initial cells	Polyclonal /R&D Systems	AF1924

### **GATA4/GATA6 Chromatin Immunoprecipitation (ChIP) sequencing**

GATA4/6 ChIPs were performed in duplicates as previously described [324]. Briefly, approximately 5x10<sup>6</sup> cells were used for each IP. Cells were cross-linked with 1% formaldehyde for 10 minutes followed by quenching with 125 mM glycine for 4-5 minutes at room temperature. The cell pellet was lysed in cell lysis buffer (20 mM Tris-HCl pH 8, 85 mM KCl, 0.5% NP-40) supplemented with 1X protease inhibitors (Roche, 11836170001) on ice for 20 minutes then spun at 5000 rpm for 10 minutes. The nuclear pellet was resuspended in sonication buffer (10 mM Tris pH 7.5, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, and 1X protease inhibitors) and incubated for 10 minutes at 4°C. In order to achieve a 200-700 bp DNA fragmentation range, nuclei were sonicated using a Bronson sonifier (model 250) with the following conditions: amplitude = 15%, time interval = 3min (total of 8-12 minutes) and pulse ON/OFF = 0.7 s/1.3 s. Chromatin was pre-cleared with Dynabeads Protein A (Invitrogen, 10002D) for 1 hour and incubated with antibody on a rotating wheel overnight at 4°C. On the following day, 30-40 µl of Dynabeads Protein A was added to chromatin for 2-3 hours. The captured immuno-complexes were washed as follows – 1x in low-salt buffer, 1x in high-salt buffer, 1x in LiCl salt buffer, and 1x in TE. The immuno-complexes were eluted in ChIP-DNA elution buffer (10 mM Tris-HCl pH 8, 100 mM NaCl, 20 mM EDTA, and 1% SDS) for 20

minutes. The eluted ChIP-DNA was reverse cross-linked overnight at 65°C, followed by proteinase K (Thermo, 25530049) treatment, RNase A (Thermo, ENO531) treatment, and Phenol:Chloroform:Isoamyl alcohol extraction. The Illumina library construction steps were carried out with 5-10 ng of purified DNA. During library construction, purification was performed after every step using QIAquick PCR purification kit (QIAGEN, 28104) or QIAquick gel extraction kit (QIAGEN, 28706). The library reaction steps were as follows: end-repair, 3' end A-base addition, adaptor ligation, and PCR amplification. The amplified libraries were size-selected for 200-450 bp on a 2% agarose E-gel (Thermo, G402002) and sequenced (single-end, 75) on a NextSeq500 or Hi-Seq2000 platform.

### 7.1.3 Cell culture and generation of transgenic/targeted cell lines

#### hiPS cell culture

ZIP13K2[211] hiPSCs were maintained in mTeSR1 (Stemcell Technologies, 85850) on pre-coated culture ware (1:100 diluted Matrigel (Corning, 354234) in KnockOut DMEM (Thermo Fisher Scientific, 10829-018)). Clump-based cell splitting was performed by incubating the cells in final 5 mM EDTA pH 8,0 (Thermo Fisher Scientific, 15575-038) in DPBS (Thermo Fisher Scientific, 14190250) 5 min at 37°C, 5% CO<sub>2</sub>. Single cell splitting was performed by incubating the cells with Accutase (Sigma-Aldrich, A6964) supplemented with 10 µM Y-27632 (Tocris, 1254) for 15 min at 37°C, 5% CO<sub>2</sub>. Cell counting was performed using a 1:1 diluted single-cell suspensions in 0,4% Trypan Blue staining-solution (Thermo Fisher Scientific, 15250061) on the Countess II automated cell-counter (Thermo Fisher Scientific). Wash-steps were performed by spinning cell-suspensions at 300 x g 5 min at room temperature (RT).

#### Generation of SOX17 CTCF loop domain knock-out hiPSC lines

ZIP13K2 hiPSCs (s. *hiPS cell culture*) [ENREF 25](#) were treated with Accutase (Sigma-Aldrich, A6964), supplemented by 10 µM Y-27632 (Tocris, 1254) for 15 min at 37°C, 5% CO<sub>2</sub> to obtain single cells. To quench and wash the cells, equal volumes of mTeSR1 were added and cells spun down for 5 min at 300 x g, 21°C. Cells were further seeded in mTeSR1 containing 10 µM Y-27632 at a density of 3 x 10<sup>5</sup> /cm<sup>2</sup> on Matrigel (Corning) pre-coated 6-well plates (Corning) and cultured 16-24 h at 37°C, 5% CO<sub>2</sub> before transfection. Transfection was carried out with up to 5 µg of modified P2X458 (including both respective sgRNAs) using Lipofectamine 3000 (Thermo Fischer Scientific) according to the manufacturers protocol. GFP<sup>+</sup> hiPSCs were FACS-sorted 16-24h post-transfection on the FACS Aria II (Beckton Dickinson) and seeded in low density (5-10x10<sup>5</sup>/55 cm<sup>2</sup>) using mTeSR1 supplemented with 10 µM Y-27632 (Tocris, 1254) to derive isogenic clones. Single cell derived colonies were picked, and half kept for maintenance respectively used for genotyping with the Phire Animal Tissue Direct PCR Kit (Thermo Fischer Scientific) accordingly. Genotypes were verified by cloning QIAquick Gel Extraction Kit (Quiagen) purified PCR products into the pJET1.2 backbone (Thermo Fischer

Scientific) and sanger sequencing of PCR single-products ([find primer sequences listed below](#)) was performed with at least 10x positively transformed 10-beta *E. coli* (NEB, C3019H) colonies.

Oligonucleotide Name	Purpose /Region	hg19 Coordinates	5'-3' Sequence
	SOX17		
SOX17_spanbound_fwd	Boundary 2 genotyping /boundary spanning	chr8:55077369-55077388	GCTCTGCACGTGGTAAAA GA
	SOX17		
SOX17_spanbound_rev	Boundary 2 genotyping /boundary spanning	chr8:55083240-55083259	TGAAGAGGACCATGAGC ACA
	SOX17		
SOX17_int_fwd	Boundary 2 genotyping /within boundary	chr8:55079982-55080001	ACACGCTAAGCCACAATG AG
	SOX17		
SOX17_int_rev	Boundary 2 genotyping /within boundary	chr8:55080374-55080393	TTCTTCACAACCTTGCCA GC
pJet1.2_fwd	Sanger sequencing	n.a.	CGACTCACTATAGGGAGA GCGGC
pJet1.2_rev	Sanger sequencing	n.a.	AAGAACATCGATTTTCCA TGGCAG

### Generation of the polyclonal SOX17-TagBFP cell line and rescue of endogenous SOX17 protein

PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA rescue construct was generated by Gibson Assembly® (NEB, E2621L) of BstBI /BamHI double-digested PB-CAG-BGHpA (Addgene Plasmid #92161) and EcoRI digested synthetically generated pUC19 DD-3xFLAG-SOX17-GS-TagBFP (Genewiz). PB-CAG-BGHpA was a gift from Xiaohua Shen (Addgene plasmid # 92161 ; <http://n2t.net/addgene:92161> ; RRID:Addgene\_92161)[325]. PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA rescue construct was deposited on addgene.org under ID #172226. Both, PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA and Super PiggyBac transposase expression vector (SBI, PB210PA-1) were co-transfected into SOX17<sup>Δ5<sup>'</sup>CTCF#8.2</sup> mTeSR1 (Stemcell Technologies) maintained human induced pluripotent stem cells [ENREF 25](#) harboring the SOX17 Boundary 2 deletion. Transfection was conducted using equimolar plasmid ratios in combination with Lipofectamine Stem Transfection Reagent (Thermo Fischer Scientific, STEM00003) according to the manufacturer's instructions. Transfected or untransfected cells were treated with mTeSR1 (Stemcell Technologies) containing 250 µg/ml m Hygromycin B (Carl Roth, 1287.1) for 2 weeks. TagBFP negative surviving cells were FACS-sorted on the FACS Aria Fusion (Beckton Dickinson) and seeded in low density (5-10x10<sup>5</sup>/55 cm<sup>2</sup>) using mTeSR1 supplemented with 10 µM Y-27632 (Tocris, 1254) to derive a polygenic /polyclonal SOX17 rescue cell line. To stabilize ectopic SOX17-TagBFP protein, undifferentiated iPSC or day 2 EN onwards differentiating cells were treated with 1 µM final Shield-1 (Takara, 632189) back-to-back with untreated controls before sample collection for downstream analysis.

### **Generation of SOX17 and eSOX17.2 CRISPR/Cas9 knock-out hiPSC lines**

ZIP13K2 hiPSCs (s. *hiPS cell culture*) were treated with Accutase containing final 10 µM Y-27632 for 15 min at 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cell suspensions were counted and seeded at a density of 1-2 x 10<sup>5</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with final 10 µM Y-27632. Cells were pre-cultured for 16 h at 37°C, 5% CO<sub>2</sub> prior to transfection.

Cells were then transfected with 6 µg of P2X458 using Lipofectamin Stem Transfection Reagent according to the manufacturer's instructions. GFP<sup>+</sup> cells were FACS-sorted 16-24 h post-transfection with the FACS Aria II or the FACS Aria Fusion (Beckton Dickinson) and seeded at a density of 0,5-1 x 10<sup>3</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with 10 µM Y-27632 to derive isogenic clones. Single-cell derived colonies were manually picked, and split half for maintenance in a well of a 96-well plate and half used for genotyping using the Phire Animal Tissue Direct PCR Kit (Thermo Fisher Scientific, F140WH) following manufacturer's instructions ([find primer sequences listed below](#)). Edited alleles were verified by cloning PCR-products into the pJET1.2 backbone (Thermo Fisher Scientific, K1232) according to the manufacturer's instructions, followed by bacterial transformation and sanger sequencing.

<b>Oligonucleotide Name</b>	<b>Purpose /Region</b>	<b>hg19 Coordinates</b>	<b>5'-3' Sequence</b>
	<i>SOX17</i>		
SOX17_cKO_outgeno_fwd	genotyping /gene body spanning	chr8:55370106 -55370125	GTCACCCACCACTGAA ACAC
	<i>SOX17</i>		
SOX17_cKO_outgeno_rev	genotyping /gene body spanning	chr8:55374049 -55374069	AATCCAGCCAATCATT TCAGC
	<i>SOX17</i>		
SOX17_cKO_ingeno_fwd	genotyping /gene body inside	chr8:55371078 -55371097	AACTGTTCTTTGCGAG CCTG
	<i>SOX17</i>		
SOX17_cKO_ingeno_rev	genotyping /gene body inside	chr8:55371700 -55371719	ACTTGTAGTTGGGGTG GTCC
	<i>eSOX17</i>		
eSOX17.2_KO_ingeno_fwd	genotyping /within <i>eSOX17.2</i>	chr8:55137584 -55137604	GAGGGTTGCTTTGCTG TGATG
	<i>eSOX17</i>		
eSOX17.2_KO_ingeno_rev	genotyping /within <i>eSOX17.2</i>	chr8:55137699 -55137719	CAGGTATGAAGGGAGT CAGGT
	<i>eSOX17</i>		
eSOX17.2_KO_outgeno_fwd	genotyping / <i>eSOX17.2</i> spanning	chr8:55138389 -55138408	ATGTTCTCCTCTGCTC TGCC



	eSOX17		
eSOX17.2_KO_outge	genotyping	chr8:55137172	CAATAGGAAGTGCTGG
no_rev	/eSOX17.2	-55137192	AAGGC
	spanning		
	Sanger		CGACTCACTATAGGGA
pJet1.2_fwd	sequencing	n.a.	GAGCGGC
	Sanger		AAGAACATCGATTTTC
pJet1.2_rev	sequencing	n.a.	CATGGCAG

### Generation of SOX17-reporter hiPS cell line

ZIP13K2 hiPSCs (s. *hiPS cell culture*) were treated with Accutase containing final 10  $\mu\text{M}$  Y-27632 for 15 min at 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cell suspensions were counted and seeded at a density of 1-2 x 10<sup>5</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with final 10  $\mu\text{M}$  Y-27632. Cells were pre-cultured for 16 h at 37°C, 5% CO<sub>2</sub> prior to transfection.

The following day, cells were transfected using Lipofectamin Stem Transfection Reagent in fresh mTeSR1 supplemented with final 10  $\mu\text{M}$  Y-27632 for 24 h at 37°C, 5% CO<sub>2</sub>. Transfection mixtures contained 3  $\mu\text{g}$  of T2A-H2B-mCitrine-loxP-hPGK-BSD-loxP donor plasmid and 3  $\mu\text{g}$  of PX335-SOX17 (1:1 molar ratio).

Two days post transfection, cells were selected with final 2  $\mu\text{g}/\text{ml}$  Blasticidin-S-HCl (Thermo Fisher Scientific, A1113903) for 14 days at 37°C, 5% CO<sub>2</sub>. For the derivation of isogenic reporter cell lines, single cell derived colonies were manually picked and expanded. Differentiation into EN followed by FACS analysis was used to confirm clones that were activating the reporter.

### Generation of LNC SOX17-promoter-KI hiPS cell line

ZIP13K2 SOX17-reporter (s. *Generation of SOX17-reporter hiPS cell line*) hiPSCs (s. *hiPS cell culture*) were treated with Accutase containing final 10  $\mu\text{M}$  Y-27632 for 15 min at 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cell suspensions were counted and seeded at a density of 1-2 x 10<sup>5</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with final 10  $\mu\text{M}$  Y-27632. Cells were pre-cultured for 16 h at 37°C, 5% CO<sub>2</sub> prior to transfection.

The following day, cells were transfected using Lipofectamin Stem Transfection Reagent in fresh mTeSR1 supplemented with final 10  $\mu\text{M}$  Y-27632 for 24 h at 37°C, 5% CO<sub>2</sub>. Transfection mixtures contained 3  $\mu\text{g}$  of mRuby-3xFLAG-NLS-3xSV40-poly(A)-loxP-mPGK-PuroR-loxP donor plasmid and 3  $\mu\text{g}$  of PX458-LNC SOX17-promoter (1:1 molar ratio).

Two days post transfection, cells were selected with final 2 µg/ml Puromycin-Dihydrochloride (Thermo Fisher Scientific, A1113803) for 14 days at 37°C, 5% CO<sub>2</sub>. For the derivation of isogenic reporter cell lines, single cell derived colonies were manually picked and expanded. Differentiation into EN followed by qRT-PCR analysis was used to confirm clones that were activating the reporter.

### Generation of CRISPRi hiPS cell line

ZIP13K2 hiPSCs (*s. hiPS cell culture*) were treated with Accutase containing final 10 µM Y-27632 for 15 min at 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cell suspensions were counted and seeded at a density of 1-2 x 10<sup>5</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with final 10 µM Y-27632. Cells were pre-cultured for 16 h at 37°C, 5% CO<sub>2</sub> prior to transfection.

The following day, cells were transfected using Lipofectamin Stem Transfection Reagent in fresh mTeSR1 supplemented with final 10 µM Y-27632 for 24 h at 37°C, 5% CO<sub>2</sub>. Transfection mixtures contained 2 µg of Super PiggyBac transposase expression vector (SBI, PB210PA-1) and 4 µg dCas9-KRAB-MeCP2 [228] (1:1 molar ratio). dCas9-KRAB-MeCP2 was a gift from Alejandro Chavez & George Church (Addgene plasmid # 110821 ; <http://n2t.net/addgene:110821> ; RRID:Addgene\_110821).

Two days post transfection, cells were selected with final 2 µg/ml Blasticidin-S-HCl (Thermo Fisher Scientific, A1113903) for 14 days at 37°C, 5% CO<sub>2</sub>. For the derivation of isogenic CRISPRi cell lines, single cell derived colonies were manually picked and expanded. IF stainings for Cas9 confirmed homogenous dCas9-KRAB-MeCP2 expression in the selected clones (*s. Immunofluorescence staining* for detailed experimental procedure).

### Production of lentiviral particles carrying sgRNAs

Lentiviral particles of specific sgRNA constructs have been produced in HEK-293T cells by co-transfection of 1:1:1 molar ratios pCMV-VSV-G plasmid (addgene, #8454 [326], 3,5µg), psPAX2 plasmid (addgene, #12260, 7µg) in combination with sgRNA specific variants of pU6-sgRNA EF1Alpha-puro-T2A-BFP [229] plasmid (addgene, #60955, 14µg). pCMV-VSV-G was a gift from Bob Weinberg (Addgene plasmid # 8454 ; <http://n2t.net/addgene:8454> ; RRID:Addgene\_8454). Prior to transfection, HEK-293T cells were grown on a 10 cm dish up to 70-80% confluency in HEK-media (KO-DMEM (Themro Fisher Scientific, 10829018), 10% fetal bovine serum (FBS, PAN Biotech, P30-2602), 1 x GlutaMAX Supplement, 100 U/ml Penicillin-Streptomycin (Thermo Fisher Scientific, 15140122) and final 1 x, 5,5 µM β-Mercaptoethanol (Thermo Fisher Scientific, 21985023)). For each sgRNA construct, plasmid DNA mixtures and 50 µl of LipoD293 transfection reagent (SignaGen Laboratories, SL100668) were mixed in 250 µl KO-DMEM at RT. After pipette mixing, transfection particles were incubated at RT for 15 min. Each sgRNA-specific mixture was added dropwise onto HEK-293T cultures in 10 ml HEK-media and incubated for 16h at 37°C, 5% CO<sub>2</sub>. Cell culture media was

exchanged by 10 ml fresh HEK-media the next day and culture supernatants (S/N) of the two subsequent days were then filtered (0,22 µm), collected and stored at 4°C. After the second harvesting day, S/N were supplemented with 1 x PEG-it virus precipitation solution (SBI, LV810A-1) for 24 h at 4°C. Viral particles were finally precipitated by centrifugation at 3234 x g, 4°C. Viral precipitates were resuspended in 200 µl mTeSR1 and either frozen at -80°C or immediately used for lentiviral transduction of CRISPRi hiPSCs. The entire lentivirus preparation and storage was carried out under S2-safety conditions and pre-cautions.

### **Lentiviral transduction of CRISPRi hiPSCs**

Lentiviral particles were either thawed on ice (if frozen) or directly used fresh on the day of production. For hiPS cells transduction, clump-based hiPSCs splitting was performed (s. *hiPS cell culture* for detailed experimental procedure) and dissociated clumps were supplemented with 10 µM Y-27632, 10 µg/ml Polybrene infection reagent (MerckMillipore, TR-1003-G) and 100 µl lentiviral particles preparation. Cells were then plated and cultured for 16 h at 37°C, 5% CO<sub>2</sub>. The following day, cells were washed 10 times with DPBS and given fresh mTeSR1 supplemented with 10 µM Y-27632 for 24 h at 37°C, 5% CO<sub>2</sub>.

Successfully infected cells were then selected with 2 µg/ml Puromycin Dihydrochloride (Thermo Fisher Scientific, A1113803) for 14 days at 37°C, 5% CO<sub>2</sub>. CRISPRi cell line expressing sgRNAs (sgLNCSOX17 and sgCtrl), were grown as bulk cultures, and Tag-BFP was used as a proxy for sgRNA expression prior to differentiation into the respective endodermal derivate.

#### **7.1.4 Differentiation and inhibitor assays**

##### **Three germ-layer differentiation**

To guarantee high reproducibility, constant media-quality, and mTeSR1 compatibility, respective germ-layer differentiations were exclusively performed utilizing the STEMdiff™ Trilineage Differentiation Kit media (Stemcell Technologies, 05230). Single cell suspensions of mTeSR1 maintained ZIP13K2 hiPSCs were seeded into the respective culture formats according to the required cell-number as recommended by the manufacturer's instructions. Media change using the STEMdiff Trilineage Differentiation Kit media was performed on a daily basis according to the manufacturer's instructions. Cells were then collected at required timepoints by washing the plate with DPBS before single-cell dissociation was performed with Accutase for 15 min at 37°C, 5% CO<sub>2</sub>. Single cell suspensions of definitive endoderm (EN) differentiated cells were utilized for further downstream analysis (qPCR, Western Blot, FACS etc.).

### **Embryoid body (EB) formation following ScoreCard assay**

ZIP13K2 hiPSC single cell suspensions were prepared and counted as previously described (s. *hiPS cell culture*). Next,  $1 \times 10^3$  cells/well of either sgCtrl or sgLNC SOX17 hiPSCs were seeded on a 96-well ultra-low attachment U-bottom plate (Corning, 7007) in respective cell culture media.

#### **Random EB differentiation**

Cells were seeded in 200  $\mu$ l /well of hES-media (Final DMEM-F12 (Thermo Fisher Scientific, 11320074), 20% KSR (Thermo Fisher Scientific, 10828028), 1% Penicillin /Streptomycin, 1% NEAA (Thermo Fisher Scientific, 11140050), 0,5% GlutaMAX, HEPES (Thermo Fisher Scientific, 31330038)), supplemented with final 10  $\mu$ M Y-27632. Single cell suspensions were spun at 100 x g for 1 min at RT and further cultured for 16 h at 37°C, 5% CO<sub>2</sub>. The following day 150  $\mu$ l media supernatant was carefully exchanged by 150  $\mu$ l fresh hES-media (without Y-27632). Cells were further cultured for additional 48 h at 37°C, 5% CO<sub>2</sub>. The very same media was replaced every 48h until day 9. At day 9, EBs were collected and pooled, washed once in DPBS and RNA isolated (s. *RNA isolation and cDNA synthesis*).

#### **Undifferentiated control EBs**

Cells were seeded in 200  $\mu$ l /well of mTeSR1, supplemented with final 10  $\mu$ M Y-27632. Single cell suspensions were spun at 100 x g for 1 min at RT and further cultured for 16 h at 37°C, 5% CO<sub>2</sub>. The following day 150  $\mu$ l media supernatant was carefully exchanged by 150  $\mu$ l fresh mTeSR1 media (without Y-27632). Cells were further cultured for additional 48 h 37°C, 5% CO<sub>2</sub>. At day 3, EBs were collected and pooled, washed once in DPBS and RNA isolated (s. *RNA isolation and cDNA synthesis*).

cDNA-conversion and ScoreCard assay (Thermo Fisher Scientific, A15870) has been performed according to the manufacturer's instructions.

#### **JNK inhibition experiments**

For the JNK-inhibition experiments, 1  $\mu$ M JNK inhibitor XVI (Sellekchem, S4901) final was supplemented to the media from day 3 of EN differentiation onward. The corresponding volume of DMSO was supplemented to the media of the control samples.

#### **Pancreatic progenitor (PP) differentiation**

Pancreatic progenitor (PP) differentiation was performed as previously described [247] with minor changes. Briefly, single cell suspensions of ZIP13K2 hiPSCs (s. *hiPS cell culture*) were seeded at a density of  $5 \times 10^5$  cells /cm<sup>2</sup> in mTeSR1 supplemented with 10  $\mu$ M Y-27632. After 24 h, culture medium was replaced with S1-media (Final 11,6 g/L MCDB131, Sigma Aldrich, M8537-1L; 2 mM D+-Glucose, Sigma Aldrich, G7528-250G; 2,46 g/L NaHCO<sub>3</sub>, Sigma

Aldrich, S5761-500G; 2% FAF-BSA, Proliant Biologicals, 68700-1; 1:50000 of 100 x ITS-X, Thermo Fisher Scientific, 51500056; 1 x GlutaMAX, Thermo Fisher Scientific, 35050-038; 0,25 mM ViatminC, Sigma-Aldrich, A4544-100G; 1% Pen-Strep, Thermo Fisher Scientific, 15140122) supplemented with final 100 ng/ml Activin-A (R&D Systems, 338-AC-01M) and 1,4 µg/ml CHIR99021 (Stemgent, 04-0004-10). The following 2 days, cells were cultured in S1-media supplemented with final 100 ng/ml Activin-A. Next, cells were cultured in S2-media (Final 11,6 g/L MCDB131; 2 mM D--Glucose; 1,23 g/L NaHCO<sub>3</sub>; 2% FAF-BSA; 1:50000 of 100 x ITS-X; 1 x GlutaMAX; 0,25 mM ViatminC; 1% Pen-Strep) supplemented with final 50 ng/ml KGF (Peprotech, 100-19-1MG) for 48 h. After these 48 h, cells were cultured in S3-media (Final 11,6 g/L MCDB131; 2 mM D--Glucose; 1,23 g/L NaHCO<sub>3</sub>; 2% FAF-BSA; 1:200 of 100 x ITS-X; 1 x GlutaMAX; 0,25 mM ViatminC; 1% Pen-Strep) supplemented with final 50 ng/ml KGF (Peprotech, 100-19-1MG), 200 nM LDN193189 (Sigma Aldrich, SML0559-5MG), 0,25 µM Sant-1 (Sigma Aldrich, S4572-5MG), 2 µM Retinoic Acid (Sigma Aldrich, R2625-50MG), 500 nM PDBU (Merck Millipore, 524390-5MG) and 10 µM Y-27632 for 24 h. Finally, cells were cultured in the previous S3-media composition w/o supplementation of LDN193189 for 24 h. Between daily media changes, cells were washed once with 1 x DPBS. Throughout the entire differentiation process, cells were cultured at 37°C, 5% CO<sub>2</sub> in 100 µl media /cm<sup>2</sup>.

### 7.1.5 Imaging/FACS based assays

#### Immunofluorescence and FACS staining – SOX17 CTCF loop domain perturbations studies

Undifferentiated (s. *hiPS cell culture*) or differentiated (s. *Three germ-layer differentiation*) ZIP13K2 cultures were treated with Accutase (Sigma-Aldrich, A6964) to obtain single cells. To quench and wash the cells, suspensions were supplemented with FACS-buffer containing final 5 mM EDTA (ThermoFischer Scientific, 15575020), 10% fetal bovine serum (FBS) (ThermoFischer Scientific, 26140079) in 1x DPBS (Thermo Fisher Scientific, 14190250). Further, cells were washed and surface-stained (ECS) in FACS-buffer for 30 minutes at 4°C using antibody-dilutions according to the manufacturer's instructions ([find antibodies listed below](#)). Cells were again washed as described above, fixed, and intracellularly stained (ICS) utilizing the True-Nuclear™ Transcription Factor Buffer Set (Biolegend, 424401) according to manufacturer's instructions ([find antibodies listed below](#)). Following subsequent wash-steps in permeabilization buffer, we performed flow cytometry data acquisition on the Celesta (Beckton Dickinson, IC-Nr.: 68186, Serial-Nr.: R66034500035). Raw data were analyzed by the use of FlowJo (Beckton Dickinson) v10.7.2.

<b>Antibody name</b>	<b>Application /Target</b>	<b>Clone /Company</b>	<b>Catalogue number</b>
Mouse anti-Human CXCR4 (CD184) PE	ECS-FACS /CXCR4	12G5 /Biolegend	306506
Mouse IgG2ak PE Isotype control	ECS-FACS /unrelated	MOPC-173 /Biolegend	400212
Mouse anti-Human Ep-CAM (CD326) PE	ECS-FACS /EpCAM	9C4 /Biolegend	324206
Mouse IgG2bk PE Isotype control	ECS-FACS /unrelated	MPC-11 /Biolegend	400314
Mouse AlexaFluor488 SOX17	ICS-FACS /SOX17	P7-969 /BD	562205
Mouse AlexaFluor488 NANOG	ICS-FACS /NANOG	N31-355 /BD	560791
Mouse IgG1k AlexaFluor488 Isotype control	ICS-FACS /unrelated	MOPC-21 /BD	557702

Undifferentiated (s. *hiPS cell culture*) or differentiated (s. *Three germ-layer differentiation*) cultures for immunofluorescent (IF) stainings were directly fixed on the culture plates, using 4% PFA solution in DPBS for 15 minutes at 21°C. Followed by multiple wash-steps with DPBS, cultures were permeabilized in PBT-buffer containing 1% BSA (Sigma-Aldrich, A2153), 10% FBS (ThermoFischer Scientific, 26140079) and 0,3% Triton-X-100 (Sigma-Aldrich, T8787) in DPBS for 30 minutes at 21°C. Blocking was further performed in PB-buffer (PBT without Triton-X-100) for 30 minutes at 21°C. Subsequently, cultures were washed in DPBS and incubated with primary or secondary antibody solutions (find antibodies below) for at least 2 hours at 21°C. DNA staining was performed using 0.25 µg/ml DAPI solution (ThermoFischer Scientific, D1306) for 15 minutes at 21°C. Microscopy was performed using the Z1 Observer (Zeiss) and fluorescent raw signals were adjusted according to the respective controls using ZEN 2 blue (Zeiss) V2.3.

<b>Antibody name</b>	<b>Application /Dilution used</b>	<b>Clone /Company</b>	<b>Catalogue number</b>
Rabbit anti-Human NANOG (unconjugated)	1 <sup>st</sup> IF /1:1000	EPR2027(2) /Abcam	ab109250
Goat anti-Human SOX17 (unconjugated)	1 <sup>st</sup> IF /1:1000	Polyclonal /R&D Systems	AF1924
Donkey anti-Rabbit IgG AlexaFluor488	2 <sup>nd</sup> IF /1:700	Polyclonal /Thermo Fisher Scientific	A21206
Donkey anti-Goat IgG AlexaFluor594	2 <sup>nd</sup> IF /1:700	Polyclonal /Thermo Fisher Scientific	A11058

### **Immunofluorescence staining – LNC SOX17 repression studies**

For immunofluorescent stainings, cells were grown in Ibidi 8-well glass-bottom plates (Ibidi, 80827) (initial seeding,  $10^4$  cells /well). On the day of analysis, cells were washed twice with DPBS and then fixed in 4% Paraformaldehyde (PFA) solution (Sigma-Aldrich, P6148-500G) for 30 min at 4°C, and then washed three more times with DPBS. Subsequently, cells were permeabilized for 30 min in DPBS-T solution (Final 0,5% Triton-X (Sigma-Aldrich, T8787-50 ML) in DPBS) and blocked for 30 min in Blocking solution (Final 10% fetal bovine serum in DPBS-T) at RT. Primary antibody incubation was performed in blocking solution for 1 h and 45 min at RT, after which cells were washed three times with Blocking solution. After the last washing step, samples were incubated with secondary antibodies diluted in Blocking solution for 30 min at RT. Afterwards, cells were washed three times with DPBS-T. The last DPBS-T washing step after secondary antibody incubation contained 0,02% DAPI (Roche Diagnostics, 10236276001). DAPI was incubated for 10 min at RT and washed off once with DPBS ([find listed antibodies and dilutions used below](#)).

<b>Antibody name</b>	<b>Application /Dilution used</b>	<b>Clone /Company</b>	<b>Catalogue number</b>
Mouse anti-Streptococcus Pyogenes Cas9	1 <sup>st</sup> IF /1:500	7A9-3A3 /SantaCruz	sc-517386

Rabbit anti-Human CXCR4	1 <sup>st</sup> IF /1:400	UMB2 /Abcam	ab124824
Rabbit anti-Human VIMENTIN	1 <sup>st</sup> IF /1:400	D21H3 /CST	5741S
Mouse anti-Human E-CADHERIN	1 <sup>st</sup> IF /1:400	36/E-Cadherin /BD	610182
Rabbit anti-Human N-CADHERIN	1 <sup>st</sup> IF /1:400	D4R1H /CST	13116T
Goat anti-Human PDX-1	1 <sup>st</sup> IF /0.5 µg per ml	Polyclonal /R&D Systems	AF2419-SP
Donkey anti-Goat IgG (H+L) Cross-Adsorbed AlexaFluor488	2 <sup>nd</sup> IF /1:700	Polyclonal /Thermo Fisher Scientific	A-11055
Donkey anti-Mouse IgG (H+L) Cross-Adsorbed AlexaFluor488	2 <sup>nd</sup> IF /1:700	Polyclonal /Thermo Fisher Scientific	A-21202
Donkey anti-Rabbit IgG (H+L) Cross-Adsorbed AlexaFluor568	2 <sup>nd</sup> IF /1:700	Polyclonal /Thermo Fisher Scientific	A-10042

### Cell clearing – *LNC5OX17* repression studies

Prior to imaging, cells were cleared with RIMS (Refractive Index Matching Solution) in order to increase light penetrability. To this end, samples were first washed three times with 0,1 M phosphate buffer (0,025 M NaH<sub>2</sub>PO<sub>4</sub>, 0,075 M Na<sub>2</sub>HPO<sub>4</sub>, pH 7,4). Clearing was then performed by incubation in RIMS solution (133% w/v Histodenz (Sigma-Aldrich, D2158) in 0,02 M phosphate buffer) at 4°C o/n.

### Immunofluorescence imaging – *LNC5OX17* repression studies

Cells stained with antibodies were imaged with the Zeiss Celldiscoverer7 (wide-field), Zeiss LSM880 (laser-scanning microscope with Airyscan), Zeiss Observer (wide-field) or Nikon Eclipse TS2 (bench-top microscope) with appropriate filters for DAPI, Alexa Fluor 488, Alexa Fluor 568, Alexa Fluor 647, and combinations thereof.



### Quantitative fluorescence microscopy – *LNC SOX17* repression studies

For each staining tested, a total of 49 individual positions were acquired in 3 fluorescence channels /replicate /well, with a 20 x /NA=0,95 objective, an afocal magnification changer 1 x , 3 x 3 camera binning, a consequential pixel size of 0,46  $\mu\text{m}^2$ , and in constant focus stabilization mode. Analysis was then performed using the Image Analysis module running in ZEN 3.2. On average 6928 single cells were analyzed per replicate. Cells were identified on smoothed nuclear counterstaining (DAPI) using fixed intensity thresholds, nearby objects were separated by mild water shedding. The consequential primary objects were filtered (area 45-175  $\mu\text{m}^2$ ) and expanded by 8 pixels (=5,44 $\mu\text{m}^2$ ); the consecutive ring, surrogated a cytoplasm compartment. Fluorescence intensities (mean and standard deviation) were quantified for each nucleus and expanded object, depending on the staining pattern profiled.

### Single molecule RNA fluorescent in situ hybridization – *LNC SOX17* repression studies

For single molecule RNA fluorescent in situ hybridization (smRNA-FISH), cells were grown in Ibidi 8-well glass-bottom plates (Ibidi 80827) (initial seeding,  $10^4$  cells /well). On the day of analysis, cells were washed twice with DPBS, fixed in 4% PFA for 10 min at RT, and washed again twice with DPBS. Cells were then incubated in 70% ethanol at 4°C for at least 1 h and then washed with 1 ml of Wash Buffer A (LGC Biosearch Technologies) at room temperature for 5 min. Cells were subsequently hybridized with 100  $\mu\text{l}$  of Hybridization Buffer (LGC Biosearch Technologies) containing the smRNA-FISH probes at a 1:100 dilution in a humid chamber at 37°C o/n (not more than 16 h). The next day, cells were washed with 1 ml of Wash Buffer A at 37°C for 30 min and stained with Wash Buffer A containing 10  $\mu\text{g}/\text{ml}$  Hoechst 33342 at 37°C for 30 min. Cells were then washed with 1 ml of Wash Buffer B (LGC Biosearch Technologies) at RT for 5 min, mounted with ProLong Gold (Thermo, P10144), and left to curate at 4°C o/n before proceeding to image acquisition. Oligonucleotides probes were designed with the Stellaris smRNA-FISH probe designer (LGC Biosearch Technologies, version 4.2), labeled with Quasar 570 and produced by LGC Biosearch Technologies ([find smRNA-FISH probe sequences listed below](#)).

Probe number	Probe sequence 5'-3'	Probe name
1	tgagaggaaccatcttcagg	lnc-SOX17_exon_Q570_1
2	cagatggaggagcctgtaaa	lnc-SOX17_exon_Q570_2
3	taagagcatcttccatgtgt	lnc-SOX17_exon_Q570_3
4	cagttgaagttggcctttat	lnc-SOX17_exon_Q570_4

5	ccaggcttatgtacagcaaa	Inc-SOX17_exon_Q570_5
6	caagaaaccttgtagccat	Inc-SOX17_exon_Q570_6
7	caatcctgggagacaaatg	Inc-SOX17_exon_Q570_7
8	cttcttagtaactgtctcca	Inc-SOX17_exon_Q570_8
9	tgctcagtagaaaacaccca	Inc-SOX17_exon_Q570_9
10	ctgcaagtacttagacacct	Inc-SOX17_exon_Q570_10
11	ccagacataggagtactgt	Inc-SOX17_exon_Q570_11
12	cagttacactttatgggctc	Inc-SOX17_exon_Q570_12
13	aagcagcatgatcagagcta	Inc-SOX17_exon_Q570_13
14	ctcttgtaattcttagtgc	Inc-SOX17_exon_Q570_14
15	agagtattgtctcttggt	Inc-SOX17_exon_Q570_15
16	gcatagatctgctagttcac	Inc-SOX17_exon_Q570_16
17	ggtggaaaacagagacccat	Inc-SOX17_exon_Q570_17
18	aactctgccggtaaaggatg	Inc-SOX17_exon_Q570_18
19	cttttcctaaggatccttt	Inc-SOX17_exon_Q570_19
20	gggcctgaattaagtgat	Inc-SOX17_exon_Q570_20
21	tagaccaggtgctatcttac	Inc-SOX17_exon_Q570_21
22	atttacacctgggagtgac	Inc-SOX17_exon_Q570_22
23	caagatgacccttgcaacat	Inc-SOX17_exon_Q570_23
24	gaatcgaaacagctgtggct	Inc-SOX17_exon_Q570_24
25	tctattgctatgtgcaatcc	Inc-SOX17_exon_Q570_25
26	gaatcttaggtcagtcctca	Inc-SOX17_exon_Q570_26
27	atatttatgctcacccttc	Inc-SOX17_exon_Q570_27

28	tcatttccttcaaccatcta	lnc-SOX17_exon_Q570_28
29	agtgattccatctccatatt	lnc-SOX17_exon_Q570_29
30	caaagtaggcagggttttca	lnc-SOX17_exon_Q570_30
31	gaggctcgagaagctgtgag	lnc-SOX17_exon_Q570_31
32	tgacattctgttgagaggg	lnc-SOX17_exon_Q570_32
33	agggagaaatgttccagctg	lnc-SOX17_exon_Q570_33
34	tgagcactcttgatagagcg	lnc-SOX17_exon_Q570_34
35	aacagcatgaaagcctgtgt	lnc-SOX17_exon_Q570_35
36	ttggcaaactctagggtttc	lnc-SOX17_exon_Q570_36
37	ggaactgtgtctttcagga	lnc-SOX17_exon_Q570_37
38	ctcattgtcaactcctcata	lnc-SOX17_exon_Q570_38
39	gtatatccttcttctgggaa	lnc-SOX17_exon_Q570_39
40	tcagtactgcgatagagtct	lnc-SOX17_exon_Q570_40
41	tggtagagaggagacctgaag	lnc-SOX17_exon_Q570_41
42	gtttaccatcttgcacatac	lnc-SOX17_exon_Q570_42
43	tggtatgctgtatctttctc	lnc-SOX17_exon_Q570_43
44	gggtagttccaaggacaatt	lnc-SOX17_exon_Q570_44
45	gaatgtgccagctacaacag	lnc-SOX17_exon_Q570_45
46	ttcagacacttcattgtct	lnc-SOX17_exon_Q570_46
47	ggaacatggtattaccctt	lnc-SOX17_exon_Q570_47
48	tgtgtttattcaagagccgt	lnc-SOX17_exon_Q570_48

### smRNA-FISH imaging – *LNC*SOX17 repression studies

Image acquisition was performed using a DeltaVision Elite widefield microscope with an Olympus UPlanSApo 100 x /1,40-numerical aperture oil objective lens and a PCO Edge

sCMOS camera. Z-stacks of 200 nm step size capturing the entire cell were acquired. Images were deconvolved with the built-in DeltaVision SoftWoRx Imaging software and maximum intensity projections were created using ImageJ [327] and Fiji [328].

### FACS staining – LNC6SOX17 repression studies

Undifferentiated or differentiated ZIP13K2 cultures were treated with Accutase for 15 min, 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. To quench the dissociation reaction and to wash the cells, FACS-buffer was added (Final DPBS, 5 mM EDTA (ThermoFisher Scientific, 15575020), 10% Fetal bovine serum (FBS, PAN Biotech, P30-2602)). Next, cells were spun down at 300 x g, 5 min at 4°C. Cells were then resuspended in FACS-buffer containing surface marker antibodies (find antibodies below) and incubated for 15 min at 4°C in the dark. For extracellular stainings (ECS) only, cells were further washed once with FACS-buffer and spun down at 300 x g before FACS analysis was performed. If additional intracellular stainings (ECS+ICS) were performed, cells were washed once with FACS-buffer, supernatants were removed and cells fixed according to the manufacturer's instructions utilizing the True-Nuclear™ Transcription Factor Buffer Set (Biolegend, 424401). Intracellular staining was performed according to manufacturer's instructions before FACS analysis was carried out. ICS antibody dilutions are listed in (find listed antibodies and dilutions used below). FACS analysis was performed on the FACSCelesta Flow Cytometer (Beckton Dickinson). Raw data were analyzed using FlowJo (LLC) V10.6.2.

Antibody name	Application /Target	Clone /Company	Catalogue number
Mouse anti-Human CRCX4 (CD184) PE	ECS-FACS /CXCR4	12G5 /Biolegend	306506
Mouse IgG2 $\alpha$ PE Isotype control	ECS-FACS /unrelated	MOPC-173 /Biolegend	400212
Mouse anti-Human PDGFR $\beta$ (CD140b) PE	ECS-FACS /PDGFR $\beta$	18A2 /Biolegend	323606
Mouse IgG1 $\kappa$ PE Isotype control	ECS-FACS /unrelated	MOPC-21 /Biolegend	400113
Mouse anti-Human NCAM (CD56) PE-Cy5	ECS-FACS /NCAM	MEM-188 /Biolegend	304608

Mouse IgG2 $\alpha$ PE-Cy5 Isotype control	ECS-FACS /unrelated	MOPC-173 /Biolegend	400218
Mouse AlexaFluor488 SOX17	ICS-FACS /SOX17	P7-969 /BD	562205
Mouse IgG1 $\kappa$ AlexaFluor488 Isotype control	ICS-FACS /unrelated	MOPC-21 /BD	557702

### 7.1.6 PCR based assays

#### RNA isolation and cDNA synthesis

For RNA extraction, cells were lysed in 500  $\mu$ l Qiazol from the miRNeasy Mini Kit (Quiagen, 217004), followed by vortexing. RNA was then extracted using the miRNeasy Mini Kit (Quiagen, 217004) and RNA concentration was measured. cDNA synthesis was performed using 1  $\mu$ g total RNA for each sample using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, K1622), following the manufacturer's instructions. Random hexamers have been used as primers for first strand cDNA synthesis.

#### Quantitative real time PCR (qRT-PCR) – SOX17 CTCF loop domain perturbations studies

TaqMan-based qRT-PCR reactions were set up in triplicate using the 2x TaqMan Fast Advanced Master Mix (Thermo, 4444557) according to manufacturer's instructions. Reactions were run on a StepOnePlus (Thermo) PCR machine with 40 cycles of 1 seconds at 95°C and 20 seconds at 60°C. TaqMan probes (Thermo) were used: *SOX17* Hs00751752\_s1; *NANOG* Hs02387400\_g1; *T/BRACHYURY* Hs00610080\_m1; *GATA4* Hs00171403\_m1; *18s* Hs03003631\_g1. *ATP6V1H* Hs00977530\_m1; *RGS20* Hs00991569\_m1; *TCEA1* Hs04403253\_g1; *LYPLA1* Hs00911024\_g1; *MRPL15* Hs00204356\_m1; *RP1* Hs00196698\_m1; *DKK1* Hs00196698\_m1; *DKK4* Hs00205290\_m1.

#### ChIP qRT-PCR – SOX17 CTCF loop domain perturbations studies

For CTCF ChIP qRT-PCR, undifferentiated ZIP13K2 cells were grown to a final count of 10 million, treated with Accutase (Sigma-Aldrich, A6964), resuspended, and washed in DPBS. Subsequently, cells were crosslinked in 1% formaldehyde solution for 5 minutes at room temperature. Following quenching with 0,125 M glycine final and two DPBS washes, we isolated nuclei using 1 ml cell lysis buffer (20 mM Tris-HCl pH8.0, 85 mM KCl, 0.5% NP 40) for 10 minutes on ice. Then nuclei were spun down for 3 minutes at 2500 x g and supernatant was removed. The pellet was resuspended in 1 ml of nuclear lysis buffer (10 mM Tris-HCl, pH 7.5, 1% NP-40, 0.5% sodiumdeoxycholate, 0.1% SDS) then incubated for 10 minutes on ice. Sonication was carried out on a Covaris E220 Evolution sonicator (PIP = 140.0, Duty Factor =

5.0, Cycles/Burst = 200, 10 minutes). After sonication, chromatin was spun down at 15000 x g for 10 minutes to pellet insoluble material. Volume was increased to 1,5 mL with Chip Dilution Buffer (0.01%SDS, 1.1% Triton X-100,1.2mM EDTA, 16.7mM Tris-HCl pH 8.1, 167mM NaCl) and 20 µl of CTCF antibody (CST, D31H2-XP) was added. Immunoprecipitation mixture was allowed to rotate overnight at 4°C. The next day, 40 µl of Protein A Dynabeads (Thermo, 10001D) were added to the IP mixture and allowed to rotate for 4 hours at 4°C. This was followed by two washes of each: low salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl pH 8.1,150mM NaCl); high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20mM Tris, pH 8.1, 500 mM NaCl); LiCl wash buffer (0.25M LCl, 1% NP40, 1% deoxycholate, 1mM EDTA, 10mM Tris-HCl pH 8.1); and TE buffer pH 8.0 (10mM Tris-HCl, pH 8.0, 1mMEDTA pH 8.0). DNA was eluted twice using 50 µl of elution buffer (0.5 to 1% SDS and 0.1 M NaHCO<sub>3</sub>) at 65°C for 15 minutes. 16 µl of reverse crosslinking salt mixture (250 mM Tris-HCl, pH 6.5, 62.5 mM EDTA pH 8.0, 1.25 M NaCl, 5 mg/ml Proteinase K) was added and samples were allowed to incubate at 65°C overnight. DNA was purified using AMPure XP beads (Beck-man-Coulter) and treated with DNase-free RNase (Roche) for 30 min at 37°C.

qRT-PCR reactions were set up in triplicate ([find primer sequences below](#)) with the 2x PowerUp SYBR Green Master Mix (Thermo, A25742). Reactions were run on a StepOnePlus (Thermo) PCR machine with 40 cycles of 15 seconds at 95°C and 60 seconds at 60°C.

<b>Oligonucleotide Name</b>	<b>Purpose /Region</b>	<b>hg19 Coordinates</b>	<b>5'-3' Sequence</b>
CTCF_ChIPqPCR_spanbound_fwd	ChIP qRT-PCR /KO span-region	chr8:55077729 -55077748	GTGCCCTCCCCAAAACATTT
CTCF_ChIPqPCR_spanbound_rev	ChIP qRT-PCR /KO span-region	chr8:55083139 -55083158	GCCTGCTCTCAAACCTTCA
CTCF_ChIPqPCR_int_fwd	ChIP qRT-PCR /CTCF-motif 2	chr8:55082039 -55082059	TGCAGTACCACATCTTGAACA
CTCF_ChIPqPCR_int_rev	ChIP qRT-PCR	chr8:55082114 -55082133	GCAAAACAACCTTACAGCGGC

		/CTCF-motif	
		2	
CTCF_ChIPqPCR_ neg_fwd	ChIP qRT- PCR /neg. ctrl. region	chr8:55262554 -55262573	TTGAGTCCCAGAGGTTGAGG
CTCF_ChIPqPCR_ neg_rev	ChIP qRT- PCR /neg. ctrl. region	chr8:55262609 -55262628	GTCTCACTTTGTCCCCTGGG
CTCF_ChIPqPCR_ pos_fwd	ChIP qRT- PCR /pos. ctrl. region	chr8:55464418 -55464437	GCCTTCAAAGCGGGTCATTT
CTCF_ChIPqPCR_ pos_rev	ChIP qRT- PCR /pos. ctrl. region	chr8:55464477 -55464496	AACCCTCACAAACCCAGACA

#### Quantitative real time PCR (qRT-PCR) – LNC5OX17 repression studies

SYBR green based qRT-PCR reactions were set up in triplicate ([find primer sequences below](#)) with the 2x PowerUp SYBR Green Master Mix (Thermo, A25742). Reactions were run on a StepOnePlus 96-well or a QuantStudio 7 Flex 384-well Real-Time PCR System (Thermo Fisher Scientific) loading 20-25ng cDNA /well with 40 cycles of 15 seconds at 95°C and 60 seconds at 60°C.

Oligonucleotide Name	Purpose /Region	hg19 Coordinates	5'-3' Sequence
hEPST11_qPCR_fw d	qRT-PCR /hEPST11 mRNA	chr13:43474475- 43474495	CAGCAGCAAGAGCA AGAAAGA
hEPST11_qPCR_rev	qRT-PCR /hEPST11 mRNA	chr13:43469216- 43469235	GGAGTCGGTCCAGA AAAGCA
hFOXA3_qPCR_fwd	qRT-PCR /FOXA3 mRNA	chr19:46376899- 46376919	GTTTTCTCTGAAGCC CACCT

hFOXA3_qPCR_rev	qRT-PCR /FOXA3 mRNA	chr19:46376948- 46376968	ACACCCTAACCAGCC TTTTCT
hGATA3_qPCR_fwd	qRT-PCR /GATA3 mRNA	chr10:8100694- 8100694	CACCCCATCACCACC TACC
hGATA3_qPCR_rev	qRT-PCR /GATA3 mRNA	chr10:8105955- 8105974	TTCACACACTCCCTG CCTTC
hGRP_qPCR_fwd	qRT-PCR /GRP mRNA	chr18:56892860- 56892879	GAACAGAAACCACCA GCCAC
hGRP_qPCR_rev	qRT-PCR /GRP mRNA	n.a. Exon 2-3 spanning	AGAACCTTTGCCTTT TGAACCT
hHHEX_qPCR_fwd	qRT-PCR /HHEX mRNA	chr10:94454413- 94454432	CTCAATGTTGCGCCT CCCCT
hHHEX_qPCR_rev	qRT-PCR /HHEX mRNA	chr10:94454481- 94454502	TATCGCCCTCAATGT CCTTC
hKLF5_qPCR_fwd	qRT-PCR /KLF5 mRNA	chr13:73636191- 73636210	AAATCCCAGAGACCG TGCGT
hKLF5_qPCR_rev	qRT-PCR /KLF5 mRNA	chr13:73636145- 73636126	GAGGAGGGGCAGTC GTTTC
hSOX17_qPCR_fwd	qRT-PCR /SOX17 mRNA	n.a. Exon 1-2 spanning	CAAGATGCTGGGCAA GTCGT
hSOX17_qPCR_rev	qRT-PCR /SOX17 mRNA	chr8:55371698- 55371717	TTGTAGTTGGGGTGG TCCTG
hCPE_qPCR_fwd	qRT-PCR /CPE mRNA	chr4:166388932- 166388953	CGTGAATGAGAAAGA AGGTGGT
hCPE_qPCR_rev	qRT-PCR /CPE mRNA	chr4:166403439- 166403460	GATTGGCAGAAAGCA CAAAGG
hqPCR_NANOG_fw d	qRT-PCR /NANOG mRNA	chr12:7947735- 7947754	GCTGAATCCTTCCTC TCCCC



hqPCR_NANOG_rev	qRT-PCR /NANOG mRNA	chr12:7947838- 7947857	GCTCCAACCATACTC CACCC
hqPCR_18s_fwd	qRT-PCR /18s RNA	chrUn_gl000220: 110654-110673	GTAACCCGTTGAACC CCATT
hqPCR_18s_rev	qRT-PCR /18s RNA	chrUn_gl000220: 110785-110804	CCATCCAATCGGTAG TAGCG
qPCR_mRuby_fwd	qRT-PCR /mRuby mRNA	n.a.	CTCTAACCTTTCCAG GCGGT
qPCR_mRuby_rev	qRT-PCR /mRuby mRNA	n.a.	GCCATCTGTCTTGCT CTTTCGT
qPCR_LNC SOX17_fwd	qRT-PCR /LNC SOX17 RNA	chr8:55140064- 55140085	GGCAACCACCACATT TTCCTT
qPCR_LNC SOX17_rev	qRT-PCR /LNC SOX17 RNA	chr8:55140172- 55140193	ACATCTCACCTCCTA CACAGAG
qPCR_GATA4_fwd	qRT-PCR /GATA4 mRNA	chr8:11565734- 11565753	TCGTTGTTGCCGTCG TTTTC
qPCR_GATA4_rev	qRT-PCR /GATA4 mRNA	chr8:11565784- 11565803	GCTTCGGTGTCTCT CTCTC
hCXCR4_qPCR_fwd	qRT-PCR /CXCR4 mRNA	Chr2:136872596- 136872616	CTGTTGTCTGAACCC CATCCT
hCXCR4_qPCR_rev	qRT-PCR /CXCR4 mRNA	chr2:136872489- 136872508	GTCCACCTCGCTTTC CTTTG
hPDX1_qPCR_fwd	qRT-PCR /PDX1 mRNA	chr13:28499155- 28499176	ACCTTGGGACCTGTT TAGAGAA
hPDX1_qPCR_rev	qRT-PCR /PDX1 mRNA	chr13:28499279- 28499298	GGTCGCCCGAGTAA GAATGG

### 5'/3' RACE PCR experiments

5'/3' rapid amplification of cDNA ends (RACE) PCR reactions were performed utilizing the 5'/3' RACE Kit, 2<sup>nd</sup> Generation (Sigma-Aldrich, 3353621001) according to the manufacturer's instructions ([find gene specific \(SP\) primer sequences listed below](#)). RACE-PCR products were cloned into pJET1.2 backbone followed by bacterial transformation and sanger sequencing.

Oligonucleotide Name	Purpose /Region	hg19 Coordinates	5'-3' Sequence
LNC SOX17_SP3_E x1	5'RACE-PCR/ Exon 1 <i>LNC SOX17</i>	chr8:55140224- 55140243	CATCCCAGGGCCTTC TTAGT
LNC SOX17_SP5_E x2	3'RACE-PCR/ Exon 2 <i>LNC SOX17</i>	chr8:55125641- 55125661	GGCAGGAGAATCACT TGAACC
LNC SOX17_SP5_E x3	3'RACE-PCR/ Exon 3 <i>LNC SOX17</i>	chr8:55123395- 55123415	TTATCTGGGTGTGGT GGTGG

### H3K9me3 Chromatin Immunoprecipitation (ChIP) qPCR

ZIP13K2-derived EN cells ( $2 \times 10^6$  /IP) were harvested, cross-linked, washed, lysed and sonicated as described previously ([s. SOX17 ChIP sequencing](#)). ChIP for H3K9me3 was performed in triplicates utilizing the High-Sensitivity ChIP Kit (abcam, ab185913) in combination with the ChIP-grade H3K9me3 antibody (ab8898, abcam) according to the manufacturer's instructions with slight modifications. Instead of DNA column purification, phenol:chloroform extraction followed by precipitation was performed to isolate DNA ([s. SOX17 ChIP sequencing](#)). Precipitated DNA was dissolved in 200  $\mu$ l H<sub>2</sub>O.

qPCR reactions were set up utilizing the 2 x PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, A25777) containing final 250 nM forward /reverse primer ([find sequences below](#)). All samples have been measured in technical triplicates using 4  $\mu$ l diluted input or IP sample from above /reaction /replicate. qPCRs were set-up on 96-well plates (Thermo Fisher Scientific, N8010560), spun down for 1 min at 2500 x g, RT and ran on a StepOnePlus 96-well Real-Time PCR System (Thermo Fisher Scientific).

<b>Oligonucleotide Name</b>	<b>Purpose /Region</b>	<b>hg19 Coordinates</b>	<b>5'-3' Sequence</b>
H3K9me3_3_ChiPqPC R_fwd	ChIP-qPCR / <i>LNC</i> SOX17 promoter	chr8:55141022- 55141043	AAGTCTCTTCCT GTTCTCCCTC
H3K9me3_3_ChiPqPC R_rev	ChIP-qPCR / <i>LNC</i> SOX17 promoter	chr8: 55140949- 55140970	AAGCAGTGGTGT GGATTTCCGG
H3K9me3_4_ChiPqPC R_fwd	ChIP-qPCR / <i>SOX</i> 17 promoter	chr8:55370163- 55370182	AGAATGGACGCT CGGTATGT
H3K9me3_4_ChiPqPC R_rev	ChIP-qPCR / <i>SOX</i> 17 promoter	chr8:55370203- 55370222	GTCTGGGAGGG CTGATTGT
H3K9me3_6_ChIP_fwd	ChIP-qPCR / <i>ANKRD30BL</i> enhancer (+)	chr13:133030904 -133030923	CCCCATCACACC CCGTAATC
H3K9me3_6_ChIP_rev	ChIP-qPCR / <i>ANKRD30BL</i> enhancer (+)	chr13:133031022 -133031041	AGCACAAAGCCC TATTCCCT
hTfrc_Intron_qPCR_2F	ChIP-qPCR / <i>TRFC</i> intron (-)	chr3:195781945- 195781966	CAGAGCAGACAT AAAGGTGAGC
hTfrc_Intron_qPCR_2R	ChIP-qPCR / <i>TRFC</i> intron (-)	chr3:195781867- 195781887	CCAACAGGAACA CACAGGAAC
eSOX17.2_KO_ingeno _fwd	ChIP-qPCR / <i>SOX</i> 17 enhancer eSOX17.2	chr8:55137584- 55137604	GAGGGTTGCTTT GCTGTGATG
eSOX17.2_KO_ingeno _rev	ChIP-qPCR / <i>SOX</i> 17 enhancer eSOX17.2	chr8:55137699- 55137719	CAGGTATGAAGG GAGTCAGGT

### 7.1.7 Miscellaneous assay types

#### Luciferase reporter assays

ZIP13K2 hiPSCs (s. *hiPS cell culture*) were treated with Accutase containing 10 μM Y-27632 for 15 min, 37°C, 5% CO<sub>2</sub> to obtain a single cell suspension. Cell suspensions were counted and seeded at a density of 10<sup>5</sup> cells /cm<sup>2</sup> in mTeSR1 supplemented with final 10 μM Y-27632. Sixteen hours later, cells were co-transfected with 15 fmol pRL-TK (Promega, E2241) and 150 fmol of either pGL4.27[luc2P/minP/Hygro] empty vector or pGL4.27[luc2P/minP/Hygro] containing either eSOX17, eSOX17.1 or eSOX17.2 utilizing Lipofectamin Stem Transfection Reagent (Thermo Fisher Scientific, STEM00003) following the manufacturer’s instructions. Transfection was performed in mTeSR1 containing 10 μM Y-27632 for 16 h at 37°C, 5% CO<sub>2</sub>. Subsequently, endoderm differentiation was initiated (day 0) using the STEMdiff Trilineage Endoderm Differentiation media. At day 0, 2, 3 or 5 of endoderm differentiation, cells were lysed and Renilla as well as Firefly Luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega, E2920) according to the manufacturer’s instructions. Raw values ([find data below](#)) were measured on the GloMax-Multi Detection System (Promega).

	1	2	3	4	5	6	7	8	9	10	11	12
A	23751.1	24290.2	24831.3	24376.3	26282.3	27396.7	26442.6	27138.2	25566.8	24456.5	22400.6	21695.3
B	409604	404363	411640	440304	445674	435794	385269	381534	381605	406889	418189	412018
C	127335	124274	128277	124007	114414	113324	141579	143159	142984	115301	117007	120018
D												
E	41327.9	38288	38092.6	43413.1	46032.7	43589.8	40509.2	41458.4				
F	159393	164820	155007	148292	151183	153877	150212	155451				
G	338447	340449	304308	220613	285435	296417	263918	271219				
H	84769.6	84532	85494.6	88767.9	119539	84048.7	86922.7	87545.1				

	1	2	3	4	5	6	7	8	9	10	11	12
A	3264.94	3108.4	3272.44	3358.3	3603.53	3304.45	3334.46	3552.24	3198.42	3262.44	3036.37	2963.36
B	8765.15	9375.6	8426.91	9067.38	8767.15	9045.35	7980.61	7863.54	7894.54	8392.89	8703.1	8613.04
C	9565.75	8907.25	9506.71	7268.17	7390.24	7904.56	9145.43	9103.4	9499.7	6229.59	6265.61	6703.84
D												
E	7882.55	7662.6	7454.28	9597.78	9179.45	9455.66	9269.52	9974.08				
F	12370.3	12632.4	11589.5	11641.6	10976.9	11489.4	12462.4	11619.5				
G	9259.51	9535.73	8773.16	8785.16	8224.77	8703.1	8791.16	8060.68				
H	8739.13	8635.06	9015.33	8394.89	8344.88	7412.25	8214.77	8296.82				

Norm.	Average background												
empty	7.685137987												
pGL4.15	7.4757	7.81437	7.56801	6.88896	7.29956	6.29035	7.9301	7.63914	7.99357	7.497	7.45105	7.35344	
pSOX17	6.18075	6.30354	6.46096										6.311417165
pLNCsox17	13.3116	13.652	13.3281	17.0617	15.4819	14.3385	15.4808	15.7259	15.0177	18.5096	16.6745	17.9025	

Norm.	Day 5 Average background												
eSOX17	2.76881	2.80284	2.73										2.76783358
eSOX17.1	7.52127	7.16386	6.86519										7.116774865
eSOX17.2	2.03489	2.64517	2.2264										2.268817811

Norm.	Average background											
empty	4.745990354											
pGL4.27	5.24296	4.80597	5.24432	4.83234	5.01476	4.60592	4.27015	4.15681				
pSOX17	12.8851	13.1695	13.3748	12.7391	13.1729	13.3393	12.8347	13.3795				
eSOX17.1	36.2273	36.1744	34.888	33.1141	34.8269	34.0588	29.702	33.8639				
eSOX17.2	9.7	9.78839	9.48325	10.574	14.3249	11.3392	10.5813	10.6516				

Day 2	Firefly											
A	7696.43	7434.27	10380.4	11009	543497	650890	518928	604974	54569.8	59006.4	63267.7	83335.8
B	7388.24	6743.86	10352.4	11477.4	528017	643587	518906	584189	54967.7	62835.5	56072.7	82208.2
C	7380.23	6561.78	10292.3	11025	536002	643020	516357	603906	57340.5	61036.4	61404.2	
D												
E	10700.7	10624.9	10566.6	11696.5	11369.3	10910.9	12720.6	12386.3	12542.4			
F	187196	178822	181937	186280	165062	165964	215900	215109	212341			
G	146869	148671	146402	132494	128029	126243	112103	98715.9	108767			
H	79613	76625.9	77082.8	77197.6	77916.1	75557.3	72247.4	71816.8	72965.8			

Norm.	Average Background											
empty	2.151241704											
pGL4.15	2.33052	2.32581	1.95593	2.15059	2.37228	2.01402	1.99888	2.18071	2.4451	2.00575	1.98498	2.05033
pSOX17	23.7195	21.2375	21.8958	21.1211	23.8663	19.7373	22.8864	21.3331	24.2711	21.2378	23.1838	22.7288
pLNCsox17	5.70249	5.7633	7.80008	6.83779	6.08977	6.07552	7.30133	6.6247	5.67492	6.0635	7.13263	

Norm.	Day 0 Norm														
eSOX17	7.57536	7.32789	6.71618										7.7376708	6.84823	2.42717
eSOX17.1	10.8945	9.62886	9.04476										1.4886683	1.33893	1.27091
eSOX17.2	5.88001	5.32312	5.33618										2.4594353	2.34801	2.35197

Day 3												
Firefly												
Read 1												
	1	2	3	4	5	6	7	8	9	10	11	12
A	6531.75	6529.75	6337.65	4264.74	4196.72	4332.77	3390.47	3134.4	3196.42			
B	4341.51	3990.77	4280.71	3471.48	3465.53	5376.52	3670.04	3437.06	3685.27			
C	2033.58	1941.43	1924.41	2076.36	2014.50	2031.79	1691.40	1630.22	2006.17			
D												
E	4402.67	4428.8	4178.72	3664.56	3128.4	3028.38	3610.53	3226.43	3282.44			
F	2416600	2439430	2595363	2313150	2239949	2316970	1874503	1838040	1827670			
G	48514.3	47845.7	47399.9	38005.1	38097.4	42726.7	56145	52800.1	64156			
H	370133	356894	351786	858248	829965	863971	665981	684753	684009	268.003	138.001	144.001

Remilia												
Read 1												
	1	2	3	4	5	6	7	8	9	10	11	12
A	13211.2	13113	12944.8	9918.03	9127.41	9137.42	5896.43	5875.42	5535.26			
B	24494.6	23672.8	23779.2	20407.1	19459.5	19495.6	19008.8	17532.6	17820.7			
C	14110.2	14596.7	13990	18307.8	16933.8	17542.6	14885.1	15744.2	15423.7			
D												
E	16597.3	16789.6	15894.4	11695.6	11229.2	10906.9	7614.38	7452.28	7704.43			
F	10682.7	10798.8	11059	12015.9	12234.1	12534.4	11639.6	11207.1	11131.1			
G	8993.31	8703.86	10092.2	6420.01	7934.65	8159.73	9011.33	9253.51	8763.14			
H	3932.63	3856.61	3888.65	8062.67	7548.34	7315.19	6281.62	7064.05	7186.12	54.0001	84.0003	74.0002

Norm.		Average background		Norm.		Day 0 Norm	
empty							
pGL4.15	0.49441	0.49796	0.4934	0.42899	0.45979	0.47418	0.53348
pSOX17	17.7244	16.9223	18.0019	17.0111	17.8069	17.3297	18.923
pLNCaOX1	14.4121	13.3005	13.7885	11.3152	11.8963	11.582	11.1615

Norm.		Average background		Norm.		Day 0 Norm	
empty							
pGL4.27	0.24357	0.28378	0.26291	0.31247	0.2786	0.27781	0.47417
pSOX17	226.225	226.806	234.683	192.507	183.09	184.849	161.045
pSOX17.1	5.30449	4.93058	4.69989	4.51518	4.80144	5.22693	6.23049
pSOX17.2	84.1194	92.3331	98.1968	104.447	109.957	119.09	108.021

Day 5												
Firefly												
Read 1												
	1	2	3	4	5	6	7	8	9	10	11	12
A	84.0003	166.001	108	122.001	220.002	48.0001	202.002	72.0002				
B	3196.42	4916.99	3524.51	3926.53	4732.92	2352.23	5031.04	2422.24				
C	3186.42	4740.92	1256.06	1314.07	4726.52	3278.44	4718.91	3186.42				
D												
E	84.0002	70.8002	78.8002	162.001	128.001	144.001	258.003	274.003	252.003			
F	3634.54	3948.61	4046.67	9915.33	8761.15	9059.36	17388.4	16591.2	16507.2			
G	352.005	384.006	342.005	318.004	384.006	314.004	762.024	794.026	692.02			
H	18611.8	19086.9	18964.7	13637.6			15367.7	13457.4				

Remilia												
Read 1												
	1	2	3	4	5	6	7	8	9	10	11	12
A	84.0003	166.001	84.0002	114.001	162.001	62.0002	134.001	92.0004				
B	162.001	242.002	160.001	194.002	228.002	120.001	240.002	126.001				
C	116.001	484.001	242.002	260.003	340.005	112.001	454.008	174.001				
D												
E	100	84.0004	100	184.001	142.001	170.001	264.003	290.003	272.003			
F	84.0003	106	78.0002	186.001	234.002	250.003	230.002	234.002	234.002			
G	102	104	130.001	170.001	148.001	152.001	176.001	134.001	148.001			
H	472.009	474.009	396.006	420.007			420.007	390.006				

Norm.		Average background		Norm.		Day 0 Norm	
empty							
pGL4.15	1	0.9881	1.65624	1.07017	1.35803	0.77419	1.50747
pSOX17	19.7309	20.318	22.028	20.2402	20.7566	19.6018	20.5625
pLNCaOX1	27.4861	9.79509	5.19029	5.05406	13.0084	29.2537	10.3939

Norm.		Average background		Norm.		Day 0 Norm	
empty							
pGL4.27	0.64	0.74468	0.78	0.88044	0.90141	0.84706	0.97727
pSOX17	43.2882	38.3076	51.8803	48.4603	37.449	36.237	75.6448
pSOX17.1	3.45103	3.69237	2.63079	1.8706	2.59462	2.0658	4.32968
pSOX17.2	41.5406	40.9881	47.8899	32.4659		38.5653	34.5026

### Western Blot

Undifferentiated or differentiated ZIP13K2 cultures were treated with Accutase for 15 min, 37°C, 5% CO<sub>2</sub> to obtain a single suspension. Single cell suspensions were washed once with ice cold DPBS and spun down at 300 x g, 5 min at 4°C. Supernatants were removed and cell lysates generated by treatment for 30 minutes on ice with RIPA buffer (Thermo Fisher Scientific, 89900) supplemented with 1 x HALT protease inhibitor (Thermo Fisher Scientific, 87786). Lysates were spun down at 12000 x g, 10 min at 4°C and supernatants quantified for protein content using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, 23227) according to the manufacturer's instructions.

For Western Blot, 20 µg total protein extract per sample were boiled in final 1 x Laemmli Buffer (BioRad, 1610747) containing 10% 2-Mercaptoethanol (M6250, Sigma-Aldrich) for 10 min at 95°C, followed by cooling on ice for 5 min. Samples were then loaded on a NuPAGE 4-12%, Bis-Tris, 1,0 mm, Mini Protein Gel (Thermo Fisher Scientific, NP0322BOX) and ran at 200 V for 30 min in 1 x NuPAGE MOPS SDS Running Buffer (Thermo Fisher Scientific, NP0001) containing 1:400 NuPAGE Antioxidant (Thermo Fisher Scientific, NP0005). Protein transfer has been performed utilizing the iBlot 2 Starter Kit, PVDF (Thermo Fisher Scientific, IB21002S) following the manufacturer's instructions for the P0 program.

PVDF membranes containing transferred proteins were incubated in blocking buffer (1 x TBS-T (Thermo Fisher Scientific, 28360), 5% Blotting-Grade Blocker (BioRad, 1706404) for 1 h at RT. Incubation with primary antibody dilution ([find antibodies below](#)) was performed in blocking buffer at 4°C overnight. The following day, membranes were washed three times 10 min at RT with 1 x TBS-T and incubated for 2 h at RT in secondary antibody dilution in blocking buffer ([find antibodies below](#)). Next, membranes were washed three times for 10 min at RT with 1 x TBS-T and developed using the SuperSignal West Dura Extended Duration Substrate (Thermo Fisher Scientific, 34075) according to the manufacturer's instructions on the BioRad ChemiDoc XRS+ imaging system.

<b>Antibody name</b>	<b>Application /Target</b>	<b>Clone /Company</b>	<b>Catalogue number</b>
Goat anti-Human SOX17 (unconjugated)	1 <sup>st</sup> WB /1:1000	Polyclonal /R&D Systems	AF1924
Mouse anti-Human LAMIN-B (unconjugated)	1 <sup>st</sup> WB 1:1000 /LAMIN-B	B-10 /SantaCruz	sc-374015
Rabbit anti-Human GAPDH (unconjugated)	1 <sup>st</sup> WB 1:5000 /GAPDH	14C10 /CTS	2118L
Rabbit anti-Human SAPK/JNK (unconjugated)	1 <sup>st</sup> WB 1:500 /SAPK/JNK	1845 /CST	9252T
Rabbit anti-Human Phospho- SAPK/JNK (Thr183/Tyr185) (unconjugated)	1 <sup>st</sup> WB 1:500 /Phospho- SAPK/JNK (Thr183/Tyr185)	1539 /CST	9251S
Donkey Anti-Goat IgG (H+L) Peroxidase AffiniPure	2 <sup>nd</sup> WB 1:10000 /Goat IgG	Polyclonal /Jackson Immunoresearch	705-035-147
Donkey Anti-Mouse IgG (H+L) Peroxidase AffiniPure	2 <sup>nd</sup> WB 1:10000 /Mouse IgG	Polyclonal /Jackson Immunoresearch	715-035-150
Donkey Anti-Rabbit IgG (H+L) Peroxidase AffiniPure	2 <sup>nd</sup> WB 1:10000 /Rabbit IgG	Polyclonal /Jackson Immunoresearch	711-035-152

## 7.2 Computational Methods and Approaches

### 7.2.1 Parameters

Default parameters were used, if not otherwise specified, for all software and pipelines utilized in this study.

### 7.2.2 Identification of CTCF loop domains from Hi-C data

Raw Hi-C reads were mapped to the hg19 version of the human genome and preprocessed using the HiC-Pro pipeline[329] (version 2.11.0-beta) to obtain uniquely mapped deduplicated interactions. These interactions were then aggregated into 10 kb genomic bins and normalized using the calCB algorithm in HiCapp[330] (v1.0.0). The high confidence (i.e. significant,  $q < 0.01$ ) interactions in the genomic range of 30 kb – 2 Mb were identified using the Fit-Hi-C python package[199] (v1.0.1). By mapping the anchors of high confidence interactions to CTCF sites (the union of CTCF motifs and CTCF ChIP-seq peaks in the corresponding sample), we obtained CTCF-CTCF loops. We observed that some samples with low sequence depth had very few identified loops, because small counts led to a low power for interactions to pass the significance cutoff. Therefore, we applied a hard cutoff to obtain the top 10,000 CTCF-CTCF loops in these samples based on previous evidence regarding the number of these loops per cell type[109, 255, 256]. Subsequently, loops close to each other were clustered and merged to reduce redundancy (see below). These merged loops generated the final set of CTCF loop domains. Note that summits of merged loops were identified based on the Hi-C interaction significance and were used instead of the merged loops themselves to increase the resolution of anchor points. The same procedure was performed by using three other Hi-C loop detection methods with the recommended parameter settings by the original references: HiCCUPS[201] (-m 500 -r 10000 -f .1), SIP[200] (-res 10000 -fdr 0.05) and Homer[202] (-res 10000 -window 50000). The CTCF loop domains were compared across different methods and were compared to insulated neighborhoods identified by cohesin ChIA-PET data in primed human ES cells[203].

### 7.2.3 Clustering and merging of redundant loops

We designed a two-step iterative clustering algorithm to cluster and merge paired-end loops within a certain genomic range cutoff; here we used a 1 kb region of boundary overlaps. In the pre-clustering step, we ranked all loops by their chromosome position and subsequently divided them into two groups based on whether they had even or odd ranks. We then used the pairToPair command in bedtools[331] (v2.25.0) to investigate the overlaps of boundaries between any paired loops from the two sets. The loops in one set that overlapped with any loops in the other set were merged to form new loops with union boundary regions. The loops in one set having no overlaps with any loops in the other set were retained. The merged and

retained loops were used as the input for the next iteration. We iteratively applied this process  $n = 50$  times to obtain pre-clustered loops. In the complete-clustering step, we used the same strategy as in the previous step, except for searching for the overlaps between the pre-clustered loops and other loops in the same set, instead of dividing them into two loop sets. Self-pairs were excluded from the analysis. In this step, the iterations were continued until the algorithm converges and no paired-end loops can be merged anymore. This two-step procedure was able to cluster and merge a large number of redundant loops in any given genomic range cutoff in a short time period.

#### **7.2.4 Boundary anchored virtual 4C visualization of Hi-C data**

To visualize the boundary interactions of many CTCF loop domains in a Hi-C data, we used the boundary anchored virtual 4C plot. It's a simple way to visualize the interactions between one boundary to the surrounding regions of the other boundary. More specifically, the left heatmap shows the Hi-C interactions between the surrounding genomic regions of the left boundaries and the right boundaries; The right heatmap shows the Hi-C interactions between the surrounding genomic regions of the right boundaries and the left boundaries. Any Hi-C matrix-like scores derived from Hi-C data can be shown by using such plot, such as the normalized Hi-C interactions and the Fit-Hi-C p-values. Then, the heatmaps can be aggregated by the columns of the left and right heatmaps to generate a shaded line plot with the line represents the average signal across columns and the shaded area represents the standard deviation signal across columns. The shaded line plot could be used to visualize the difference between multiple groups of interactions as well as calculate statistics. The left and right heatmaps are plotted into single shaded line plot with a dotted vertical line to separate them.

#### **7.2.5 Identification of CTCF motifs and their conservation across species**

CTCF motif loci and orientations in the hg19 version of the human genome were identified using FIMO<sup>[332]</sup> (v4.11.1). For this analysis, we used the consensus CTCF motif MA0139.1 from the JASPAR CORE 2016 vertebrates database<sup>[333]</sup>. Motif conservation information was obtained from the UCSC “phastCons46wayPlacental” track.

#### **7.2.6 Evolutionary analysis of human CTCF loop domain boundaries**

The CTCF motif coordinates of human CTCF loop domain boundaries were liftovered to 45 vertebrate genomes with parameter: -minMatch=0.9. The motifs successfully liftovered were called present, otherwise absent, in the corresponding genome. The percent of present motifs in different CTCF loop domain groups across species were studied.

#### **7.2.7 Identification of consensus CTCF binding sites**

The CTCF ChIP-seq peaks in 142 different cell lines and tissues (s. Wu H.J. and Landshammer A. et al., 2021 [Supplementary Data 5](#)), which were identified using the same settings and



contained at least 10,000 peaks, were downloaded from the Cistrome database[334]. The CTCF peaks ( $p < 1e-9$ , peak significance over input) detected in more than 30% of all unique cell types were defined as consensus CTCF binding sites. The coordinates of ChIP-seq peaks were overlaid with CTCF motifs to obtain orientation information and highest resolution of CTCF binding sites. Specifically, for the ChIP-seq peaks overlapping with CTCF motif(s), the motif coordinates and orientations were used instead of the peak coordinates. For the ChIP-seq peaks not overlapping with any CTCF motif, the peak coordinates were used, and the orientations were set as 'unclear'.

### 7.2.8 Clustered and typical CTCF-binding sites

CTCF ChIP-seq data was analyzed in a similar way as the enhancer analysis of the ROSE pipeline[284]. Specifically, CTCF peaks were merged within a maximal distance of 12.5 kb. The merged peaks were ranked by increasing total ChIP-seq signal and plotted against the total ChIP-seq signal. This plot showed a clear transition point in the distribution of CTCF occupancy where the total signal began increasing rapidly. The transition point was the x axis point for which a line with a slope of 1 was tangent to the curve. We then defined peaks above this point to be clustered CTCF-binding sites, and peaks below that point to be typical CTCF-binding sites. Thus, clustered CTCF-binding sites represent those sites with broad and high CTCF occupancy, while typical CTCF-binding sites represent sites with narrow and low CTCF occupancy.

### 7.2.9 Identification of enhancers and analysis of their H3K27ac enrichment

Enhancers were collected from both the Fantom5 database[152] and the Roadmap Epigenomics project[335]. The enhancers from Fantom5 were directly downloaded from the website ([http://slidebase.binf.ku.dk/human\\_enhancers/](http://slidebase.binf.ku.dk/human_enhancers/)). The enhancers from the Roadmap Epigenomics project were identified from H3K27ac ChIP-seq data of 98 samples. The aligned reads from 98 samples were downloaded as bed files and converted to bam and bigwig files using MACS2[336] (v2.1.0.20150731) and bedtools[331] (v2.25.0). Narrow peaks ( $p < 1e-9$ , peak significance over input) in the same samples called from MACS2 were downloaded and used to identify enhancers via the ROSE pipeline[284] (v0.1). The enhancers from both databases were merged and the union sets were designated as the final list of enhancers. Average H3K27ac signals for enhancers were obtained from the ChIP-seq bigwig files and normalized to signals per 10 million reads for each library. Forty-five of 98 samples for which both H3K27ac ChIP-seq and RNA-seq data was available were used for the analysis (s. Wu H.J. and Landshammer A. et al., 2021 [Supplementary Data 5](#)). Enhancers with a maximal signal of less than 5 across all 45 samples were treated as inactive enhancers and removed from the analysis. We found that our results were robust when this cutoff was changed.

### 7.2.10 Gene sets and enrichment analysis

Developmental regulators were genes overlapping with transcription factors and genes under the GO term GO:0032502 - developmental processes. Early developmental regulators were obtained from Tsankov et al.[29]. Gene set enrichment analysis was performed using Fisher's exact test.

### 7.2.11 Chromatin confirmation capture sequencing analysis

#### Capture Hi-C (cHi-C) sequencing analysis

Raw sequence reads of capture Hi-C (cHi-C) were mapped to the hg19 version of the human genome using BWA (v0.7.17) with parameters (mem -A 1 -B 4 -E 50 -L 0). Mapped reads were further processed by HiCExplorer (v3.5.1) to remove duplicated reads and reads from dangling ends, self-circle, self-ligation, and same fragments. The replicates of the same samples were compared and confirmed to have high consistency (Pearson correlation coefficient: 0.83-0.99), then were merged to construct contact matrices of 2 kb resolution. Normalization was performed to ensure that all samples have the same number of total contacts, followed by KR normalization. The relative contact difference between two cHi-C maps was calculated by subtracting one from the other after scaling one sample to the other by using the total number of contacts in each sample.

#### Circular chromatin confirmation capture (4C) sequencing analysis

The raw sequencing reads were trimmed by using cutadapt [337] (--discard-untrimmed -e 0.05 -m 25) to remove primer sequences and restriction enzyme sequences. The reads not matching those sequences, were removed from further analysis. The remaining reads were then mapped to the reference sequences GRCh37/hg19 by bowtie2 [338] (default parameters). An iterative mapping procedure was performed. Specifically, the full-length reads were first mapped to the genome. The unmapped reads were then cut by 5-nt from the 3-prime end each time until they were successfully mapped to the genome or until they were shorter than 25 bp. The final mapped reads were assigned to valid fragments. The fragment counts were then normalized by RPM (reads per million) and smoothed by averaging the counts of the closest 5 fragments.

### 7.2.12 Coding potential calculation

Whole genome multiple species alignments of 46 vertebrate species with human (assembly hg19, October 2009) as a reference have been retrieved from the UCSC genome browser[339]. Human lincRNA annotation was obtained from Gencode[340] (gencode.v33lift37.long\_noncoding\_RNAs.gtf, December 2019). All ORFs in each transcript were identified and the corresponding multiple species alignment was scored by the omega method of PhyloCSF[227] (Fig. 25B left panel) shows 95% (2.5-97.5percentile) of the 271,572

sORFs from the [339] analyzed human lincRNAs. The *SOX17* CDS and all identified sORFs in *LNC**SOX17* were scored by omega phyloCSF as shown in (Fig. 25B right panel), right panel.

### 7.2.13 ChIP sequencing

#### SOX17 Chromatin Immunoprecipitation

The ChIP-seq sequencing data as well as the control input sequencing were aligned to the human reference genome (hg19) using BWA mem [341] using the default parameter. GATK [342] was used to obtain alignment metrics and remove duplicates. Peaks were called using the MACS2 (2.1.2\_dev) [336] peakcall function using default parameters. After validation of replicate comparability and quality, replicates were merged on read level and reprocessed together with input samples. Background subtracted coverage files were obtained using MACS2 bdgcomp with -m FE.

#### GATA4/6 Chromatin Immunoprecipitation

The ChIP-seq sequencing data as well as the Fastqs for GATA4/6 ChIP-seq experiments were processed using the ENCODE ChIP-seq pipeline version 1.6.1 (<https://github.com/ENCODE-DCC/chip-seq-pipeline2>) using default settings with the hg19 genome. Standard ENCODE ChIP-seq reference files were used as found in [https://storage.googleapis.com/encode-pipeline-genome-data/genome\\_tsv/v1/hg19\\_caper.tsv](https://storage.googleapis.com/encode-pipeline-genome-data/genome_tsv/v1/hg19_caper.tsv). Pooled fold-change bigWigs were used.

### 7.2.14 Single-cell RNAseq pipeline

Publicly available single-cell RNAseq raw data of already filtered 1195 cells from a gastrulating human embryo [17] was downloaded from ArrayExpress [343] under accession code E-MTAB-9388. The GENCODE [344] human transcriptome (GRCh37.p13) and its annotation were downloaded and added with the *LNC**SOX17* entry. After building the transcriptome index, the transcripts abundance was quantified via Salmon v1.6.0 [345] in quasi-mapping-based mode using the `--seqBias` and the `--gcBias` flags. Data was loaded as a scanpy v1.4.4 [346] object, reproducing clustering as reported by Tyser, R. C. v. et al. [17]. The resulting clusters were visualized via the scanpy UMAP representation in two dimensions, using default parameters (`tl.umap`). UMAPs are displayed in (Fig. 24B upper panel).

### 7.2.15 Bulk measurements from scRNAseq pipeline

To measure *LNC**SOX17* read counts in endoderm cells fastq files were combined in one bulk raw file. The file went through a bulk RNAseq pipeline comprising a pre-alignment quality control via fastQC v0.11.9, adaptor and low-quality bases trimming using cutadapt [337], post-QC and reads alignment against the human genome (GRCh37.p13) by means of STAR [347] (parameters: `--outSAMtype BAM SortedByCoordinate, --chimSegmentMin 20, --`

outSAMstrandField intronMotif, --quantMode GeneCounts). Finally, the BAM file was visualized using the Integrative Genomic Viewer (IGV) [348]. IGV tracks are displayed in (Fig. 24B lower panel).

### 7.2.16 Oxford Nanopore RNA analysis

All Oxford Nanopore Technologies derived runs were processed using the Nanopype pipeline (v1.1.0) [349]. The basecaller Guppy (v4.0.11) was used with the r9.4.1 high-accuracy configuration. Quality filtering was disabled for any base calling. Base-called reads were aligned against the human reference genome hg19 using minimap2 (v2.10) [350] with the Oxford Nanopore Technologies parameter preset for spliced alignments (-ax splice -uf -k14). Only unique alignments (-F 2304) are reported.

### 7.2.17 Oxford Nanopore RNA split-read analysis

Nanopore post processed split read data (s. Oxford Nanopore RNA analysis) from wild type endoderm mRNA (s. Extraction of poly(A) RNA for Nanopore sequencing; s. Preparation of Nanopore sequencing libraries) were extracted from the junctions-track of BAM files visualized using the Integrative Genomic Viewer (IGV) [348] utilizing the coordinates hg19, chr8:55115873-55141447. Split reads between hg19, chr8:55140801 (5'-sequence of Exon 1, s. 5'/3' RACE PCR experiments) and hg19, chr8: 55125601 (3'-sequence of Exon 3, s. 5'/3' RACE PCR experiments) were accounted for isoform Ex1+2 (find raw data below).

Read #	Chr	Start	End	Strand	Isoform
1	chr8	55127102	55139691	-	1+2
2	chr8	55127111	55139692	+	1+2
3	chr8	55127126	55139676	+	1+2
4	chr8	55127152	55139676	-	1+2
5	chr8	55127167	55139656	+	1+2
6	chr8	55127188	55139643	+	1+2
7	chr8	55126391	55138809	+	1+2
8	chr8	55127142	55139338	-	1+2
9	chr8	55127100	55139245	+	1+2
10	chr8	55127150	55139278	-	1+2
11	chr8	55127155	55139268	+	1+2
12	chr8	55129190	55139719	+	1+2
13	chr8	55129219	55139667	-	1+2
14	chr8	55129145	55139279	+	1+2
15	chr8	55129234	55139245	-	1+2
16	chr8	55131363	55139656	+	1+2
17	chr8	55123296	55140029	-	1+3
18	chr8	55124857	55139675	+	1+3
19	chr8	55125544	55140214	+	1+3
20	chr8	55125052	55139652	-	1+3
21	chr8	55125119	55139683	-	1+3
22	chr8	55125736	55140297	+	1+3
23	chr8	55125119	55139671	-	1+3
24	chr8	55125125	55139677	+	1+3
25	chr8	55125872	55140157	+	1+3
26	chr8	55125130	55139251	+	1+3
27	chr8	55125728	55139767	+	1+3
28	chr8	55118333	55134960	+	"Sloppy"
29	chr8	55128806	55139460	+	"Sloppy"
30	chr8	55118862	55128996	-	"Sloppy"
31	chr8	55125588	55134765	+	"Sloppy"
32	chr8	55123181	55132353	+	"Sloppy"
33	chr8	55122998	55132030	-	"Sloppy"
34	chr8	55127995	55136237	-	"Sloppy"
35	chr8	55133298	55139800	+	"Sloppy"
36	chr8	55134188	55139275	-	"Sloppy"
37	chr8	55119013	55122386	+	"Sloppy"
38	chr8	55118970	55121832	-	"Sloppy"
39	chr8	55136129	55138677	-	"Sloppy"
40	chr8	55133171	55135715	-	"Sloppy"
41	chr8	55136282	55138762	-	"Sloppy"
42	chr8	55119017	55121305	+	"Sloppy"
43	chr8	55118751	55120898	-	"Sloppy"
44	chr8	55118866	55120888	-	"Sloppy"
45	chr8	55118985	55120859	-	"Sloppy"
46	chr8	55119094	55120393	-	"Sloppy"
47	chr8	55118887	55120048	+	"Sloppy"
48	chr8	55120032	55121111	-	"Sloppy"
49	chr8	55118996	55119893	-	"Sloppy"
50	chr8	55138734	55139622	-	"Sloppy"
51	chr8	55119009	55119889	-	"Sloppy"
52	chr8	55120076	55120914	-	"Sloppy"
53	chr8	55119068	55119879	-	"Sloppy"
54	chr8	55119066	55119830	-	"Sloppy"
55	chr8	55120057	55120680	-	"Sloppy"
56	chr8	55120076	55120691	+	"Sloppy"
57	chr8	55134773	55135297	+	"Sloppy"
58	chr8	55139267	55139713	-	"Sloppy"
59	chr8	55139244	55139656	-	"Sloppy"
60	chr8	55139268	55139666	+	"Sloppy"
61	chr8	55139336	55139686	-	"Sloppy"
62	chr8	55139312	55139654	-	"Sloppy"
63	chr8	55139333	55139666	+	"Sloppy"
64	chr8	55138870	55139194	+	"Sloppy"
65	chr8	55140489	55140733	+	"Sloppy"
66	chr8	55137111	55137329	-	"Sloppy"
67	chr8	55138171	55138347	-	"Sloppy"
68	chr8	55124625	55124789	-	"Sloppy"
69	chr8	55120493	55120626	+	"Sloppy"
70	chr8	55136859	55136959	-	"Sloppy"
71	chr8	55132327	55132424	-	"Sloppy"
72	chr8	55131701	55139695	+	"Sloppy"
73	chr8	55125118	55132409	+	"Sloppy"
74	chr8	55133078	55140328	-	"Sloppy"
75	chr8	55133164	55139797	-	"Sloppy"
76	chr8	55122995	55127887	+	"Sloppy"
77	chr8	55129005	55133107	+	"Sloppy"
78	chr8	55127177	55131060	+	"Sloppy"
79	chr8	55133126	55136128	-	"Sloppy"
80	chr8	55133100	55136090	-	"Sloppy"
81	chr8	55133125	55136071	-	"Sloppy"
82	chr8	55128045	55130588	+	"Sloppy"
83	chr8	55125283	55127734	-	"Sloppy"
84	chr8	55125553	55127902	-	"Sloppy"
85	chr8	55123302	55125543	+	"Sloppy"
86	chr8	55127111	55129238	+	"Sloppy"
87	chr8	55125045	55127085	+	"Sloppy"
88	chr8	55127166	55129189	-	"Sloppy"
89	chr8	55123551	55125566	-	"Sloppy"
90	chr8	55125125	55127064	-	"Sloppy"
91	chr8	55125671	55127576	+	"Sloppy"
92	chr8	55125635	55127434	+	"Sloppy"
93	chr8	55127770	55129293	+	"Sloppy"
94	chr8	55125664	55127178	-	"Sloppy"
95	chr8	55132030	55133347	+	"Sloppy"
96	chr8	55125645	55126868	+	"Sloppy"
97	chr8	55125647	55126741	+	"Sloppy"
98	chr8	55125636	55126690	+	"Sloppy"
99	chr8	55128025	55128920	-	"Sloppy"
100	chr8	55128013	55128847	-	"Sloppy"
101	chr8	55125637	55126469	-	"Sloppy"
102	chr8	55125647	55126476	+	"Sloppy"
103	chr8	55125658	55126476	-	"Sloppy"
104	chr8	55125645	55126461	-	"Sloppy"
105	chr8	55120154	55120956	+	"Sloppy"
106	chr8	55125674	55126463	-	"Sloppy"
107	chr8	55120150	55120897	-	"Sloppy"
108	chr8	55133105	55133807	+	"Sloppy"
109	chr8	55127045	55127683	+	"Sloppy"
110	chr8	55120133	55120748	-	"Sloppy"
111	chr8	55125822	55126426	-	"Sloppy"
112	chr8	55120112	55120691	+	"Sloppy"
113	chr8	55120124	55120673	+	"Sloppy"
114	chr8	55120128	55120672	+	"Sloppy"
115	chr8	55121273	55121796	+	"Sloppy"
116	chr8	55127413	55127755	-	"Sloppy"
Ex1+2		16	13,8%		
Ex1+3		11	9,5%		
Sloppy		89	76,7%		

Split reads between hg19, chr8:55140801 (5'-sequence of Exon 1, s. 5'/3' RACE PCR experiments) and hg19, chr8:55123254 (3'-sequence of Exon 3, s. 5'/3' RACE PCR experiments) were accounted for isoform Ex1+3 (find raw data below). All other reads were accounted as "sloppy spliced" reads and together with both isoforms calculated in relative terms (find raw data below). Summary of the relative isoform quantification is displayed in (Fig. 25A).

### 7.2.18 RNA sequencing data analysis

#### **SOX17 CTCF loop domain perturbations studies**

RNAseq data were pre-processed using cutadapt[351] to remove adapter sequences and trim low-quality bases. Reads were aligned against hg19 using STAR [347] (v 2.6.1d, parameter: -outSAMtype BAM SortedByCoordinate --outSAMattributes Standard --outSAMstrandField intronMotif --outSAMunmapped Within --quantMode GeneCounts). Subsequently, Stringtie[352] (v 1.3.5) was used for transcript assembly, e.g. calculation of strand-specific TPMs. Differential expression analysis was done independently per group comparison using the R package DESeq2[353] utilizing the raw expression counts from STAR's reads per gene output and filtered for an adjusted p-value < 0.05 and a log2 foldchange > 1. The PCA was calculated on the log2+1 normalized TPMs of the 100 most variable genes using the R function prcomp (parameters "center = TRUE, scale = TRUE"). Box- and scatter plots show the unmodified TPMs. The heatmaps shows Z-score normalized TPMs to adjust for differences in absolute expression levels and was plotted using the R package pheatmap.

#### **LNC SOX17 repression studies**

All RNAseq samples were pre-processed using cutadapt [337] to remove adapter and trim low quality bases. Reads were subsequently aligned against the human reference genome hg19 using STAR [347] (parameter: outSAMtype BAM SortedByCoordinate --outSAMattributes Standard --outSAMstrandField intronMotif --outSAMunmapped Within --quantMode GeneCounts). Finally, Stringtie [352] was used for calculation of strand specific TPMs.

### 7.2.19 Data visualization

#### **SOX17 CTCF loop domain perturbations studies**

Juicebox[354] was used to generate .hic files of Hi-C data to visualize in the WashU EpiGenome Browser[355] to create the genome track figures. Other figures were plotted in the R environment (<https://www.r-project.org>) using basic plotting functions and packages of ggplot2[356], pheatmap (<https://cran.r-project.org/web/packages/pheatmap/index.html>). APA plots were generated by Juicebox[354].

### **LNC $SOX17$ repression studies**

Command-line processing of BAM, BED and bigwig files was done using SAMtools (v1.10) [357], BEDtools (v2.25.0) [331] and UCSCtools (v4) [358]. If not stated otherwise: All statistics and plots are generated using R version 3.6.0 and 3.6.1 and GraphPad Prism 8. In all boxplots, the centerline is median; boxes, first and third quartiles; whiskers, 1.5 x inter-quartile range; data beyond the end of the whiskers are displayed as points.

#### **7.2.20 Data and code availability**

##### **SOX17 CTCF loop domain perturbations studies**

All Hi-C, RNA-seq and capture Hi-C data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE127196 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127196>]. The Hi-C data used in this study are available in the GEO database under accession number GSE52457 [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52457>]. Hi-C data of human embryos were obtained from the Genome Sequence Archive with the accession number CRA000852. CTCF ChIP-seq data used in this study are available at Cistrome [<http://cistrome.org>]. RNA-seq and H3K27ac ChIP-seq data used in this study are available at Epigenomics Roadmap Project [<http://www.roadmapepigenomics.org>]. Enhancers used in the study are available at Fantom5 database [<https://fantom.gsc.riken.jp/5>] and Epigenomics Roadmap Project [<http://www.roadmapepigenomics.org>]. Other data supporting the findings including source data of this study are provided online in the paper of Wu, H.J. and Landshammer A. et al., 2021. The computational code used in the manuscript is available at [https://bitbucket.org/mthjwu/loop\\_cluster](https://bitbucket.org/mthjwu/loop_cluster).

##### **LNC $SOX17$ repression studies**

All data presented in this study are available in the main text, methods, or tables. Sequencing data have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE178990. Codes used to perform the analysis in this study are available upon request.

## 8 APPENDICES

### 8.1 Acknowledgement

First and foremost, my gratitude goes to Prof. Dr. Alexander Meissner for giving me the opportunity to work in his lab, for letting me pursue all the different interesting and fascinating projects and for his brilliant supervision and terrific guidance, enhancing these projects and improving my way to do science.

I am also greatly thankful to my thesis advisors Prof. Dr. Katja Nowick and Dr. Daniel Ibrahim, for their fabulous help, insights, and suggestions, improving my conducted work to the highest level possible.

I want to thank all the members of the Meissner lab, respectively the whole Department for Genome Regulation creating a great and including lab-environment for everyone.

I would like to give many special thanks to my great friend Adriano Bolondi, who has always been an amazing lab-buddy, a great and outstanding collaborator, scientist, and the best private Italian chef I ever met. Thank you man, I am sure you will have a great carrier in front of you as a big scientist and scientific visionary!

I want to express many thanks to Christina Riemenschneider, who has been an amazingly funny and entertaining car-driving company over all the years. Thank you for listening to my silly stories every morning, you will make your way as a very thoughtful and great scientist.

A big “thank you” must be shouted out for Abhishek Sampath Kumar, an outstanding scientist with an awesome personality, always seeing the glass half-full and being a great friend even in the worst times of Ph.D. and life. Thank you!

I also want to thank, Dr. Helene Kretzmer for her amazing computational support and great chats about “god and the world”. Thank you so much, without you my projects would have never been what they turned out to be at the end. Hope to see you having your own lab soon!

I am greatly indebted to all my friends and family who always supported me during all these years in good and bad times, thank you so much! Especially my brother Michael Landshammer and my mother Christine Neun, who went through the worst times in life, I am so sorry I couldn't be there to support you more than required. I will never forget what you have been doing for our family, my father Sigfried Landshammer, my great-grandma Anni Regelin and me. Thank you forever!

I want to express my very last and most thanks to Maike Marczenke, who has always been a great person, lovely partner, and a great scientist all over the last years. I hope you may find your way back to the bench, science definitely needs you!

## 8.2 Declaration of Academic Integrity

I hereby confirm that this thesis on

*“Study of the human SOX17 locus and its genetic determinants in definitive endoderm”*

is solely my own work and that I have used no sources or aids other than ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

13.04.2022, Alexandro Landshammer

---

(date and signature of student)

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

13.04.2022, Alexandro Landshammer

---

(date and signature of student)



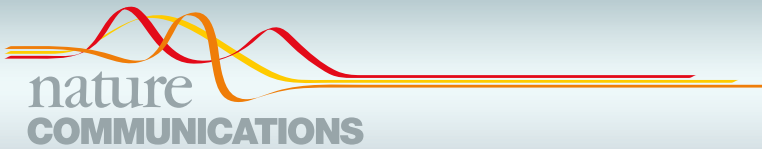
### **8.3 Curriculum Vitae**

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

## 8.4 Attachment






## ARTICLE


<https://doi.org/10.1038/s41467-021-24951-7>

OPEN

# Topological isolation of developmental regulators in mammalian genomes

Hua-Jun Wu<sup>1,2,3,12</sup>, Alexandro Landshammer<sup>4,5,12</sup> , Elena K. Stamenova<sup>6</sup>, Adriano Bolondi<sup>4,5</sup>, Helene Kretzmer<sup>4</sup>, Alexander Meissner<sup>4,5,6,7</sup>  & Franziska Michor<sup>6,7,8,9,10,11</sup> 

Precise control of mammalian gene expression is facilitated through epigenetic mechanisms and nuclear organization. In particular, insulated chromosome structures are important for regulatory control, but the phenotypic consequences of their boundary disruption on developmental processes are complex and remain insufficiently understood. Here, we generated deeply sequenced Hi-C data for human pluripotent stem cells (hPSCs) that allowed us to identify CTCF loop domains that have highly conserved boundary CTCF sites and show a notable enrichment of individual developmental regulators. Importantly, perturbation of such a boundary in hPSCs interfered with proper differentiation through deregulated distal enhancer-promoter activity. Finally, we found that germline variations affecting such boundaries are subject to purifying selection and are underrepresented in the human population. Taken together, our findings highlight the importance of developmental gene isolation through chromosomal folding structures as a mechanism to ensure their proper expression.

<sup>1</sup>Center for Precision Medicine Multi-Omics Research, Peking University Health Science Center, Beijing, China. <sup>2</sup>School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China. <sup>3</sup>Peking University Cancer Hospital and Institute, Beijing, China. <sup>4</sup>Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany. <sup>5</sup>Institute of Chemistry and Biochemistry, Freie Universität Berlin, Berlin, Germany. <sup>6</sup>The Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>7</sup>Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. <sup>8</sup>Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>9</sup>Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. <sup>10</sup>The Ludwig Center at Harvard, Boston, MA, USA. <sup>11</sup>Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>12</sup>These authors contributed equally: Hua-Jun Wu, Alexandro Landshammer. ✉email: [meissner@molgen.mpg.de](mailto:meissner@molgen.mpg.de); [michor@jimmy.harvard.edu](mailto:michor@jimmy.harvard.edu)

Mammalian genomes are organized into a hierarchy of local structures including megabase-sized topologically associating domains (TADs) and DNA loops that are usually localized within TADs<sup>1–10</sup>. The majority of TADs are stable across cell types and conserved between mouse and human<sup>2,8</sup>, while DNA loops of enhancer-gene interactions are generally more celltype-specific<sup>11,12</sup>. Recent studies identified specific DNA loops organized by CCCTC-binding factor (CTCF) and cohesin; these loops, called insulated neighborhoods, are local structures within TADs that encompass most enhancer-promoter loops<sup>5</sup>.

Disrupting the boundaries of TADs or insulated neighborhoods can lead to novel chromosomal interactions and ectopic long-range enhancer adoption, which can interrupt key gene function<sup>3,13</sup>. Such altered boundary elements, usually caused by structural variations, can also lead to developmental disorders in humans. The first human to mouse translational case study focused on abnormal limb syndromes caused by genomic alterations at the TAD boundaries containing the *EPHA4* locus. In this study, a cluster of limb enhancers normally associated with the *Epha4* gene was found to be misplaced and to ectopically activate genes, including *Wnt6*, *Pax3*, and *Ihh*, in the neighboring TADs<sup>14</sup>. A related study showed that genomic duplication of a murine boundary between *Kcnj2* and *Sox9* TADs resulted in the formation of new TADs, the ectopic activation of *Kcnj2* and the onset of Cooks Syndrome—another limb malformation<sup>15</sup>. However, duplication of smaller DNA segments at the same locus within the *SOX9* TAD causes a different phenotype, that of sex reversal, in humans<sup>16</sup>. Moreover, different chromosomal conformations of the *Pitx1* locus have been shown to lead to activation of *Pitx1* by ectopic interactions with its active enhancer *Pen* in the forelimb, causing partial arm-to-leg transformation in both human and mouse<sup>17</sup>. In another example, a large genomic deletion leading to enhancer adoption by the *LMNB1* gene was identified as an alternative path to autosomal dominant adult-onset demyelinating leukodystrophy<sup>18</sup>. In addition, in 273 subjects with congenital anomalies, 7.3% of balanced chromosomal abnormalities (BCAs) disrupted TADs containing known syndromic loci; for instance, breakpoints of BCAs in eight subjects disrupted the *MEF2C*-containing TAD, resulting in decreased expression of *MEF2C*, which is linked to 5q14.3 microdeletion syndrome<sup>19</sup>.

These selected case studies suggest a crucial role for insulated chromosome structures in gene regulation, raising the question whether this is a more universal mechanism that contributes to precise gene control by limiting domain-level access to regulatory elements in development. One previous study demonstrated gene expression changes upon boundary disruptions in mouse ES cells<sup>5</sup>, but it remains incompletely understood how insulation boundaries influence early stem and progenitor differentiation. Prior work has also demonstrated higher sequence conservation across primates<sup>20</sup> and elevated somatic mutation rates across tumor types<sup>10</sup> in the boundary CTCF motifs of insulation domains as compared to other sequences; however, it remains to be shown whether and how insulating structures shape the gene distribution across the human genome and whether there are associations between the function of an insulating domain and the number of genes it contains.

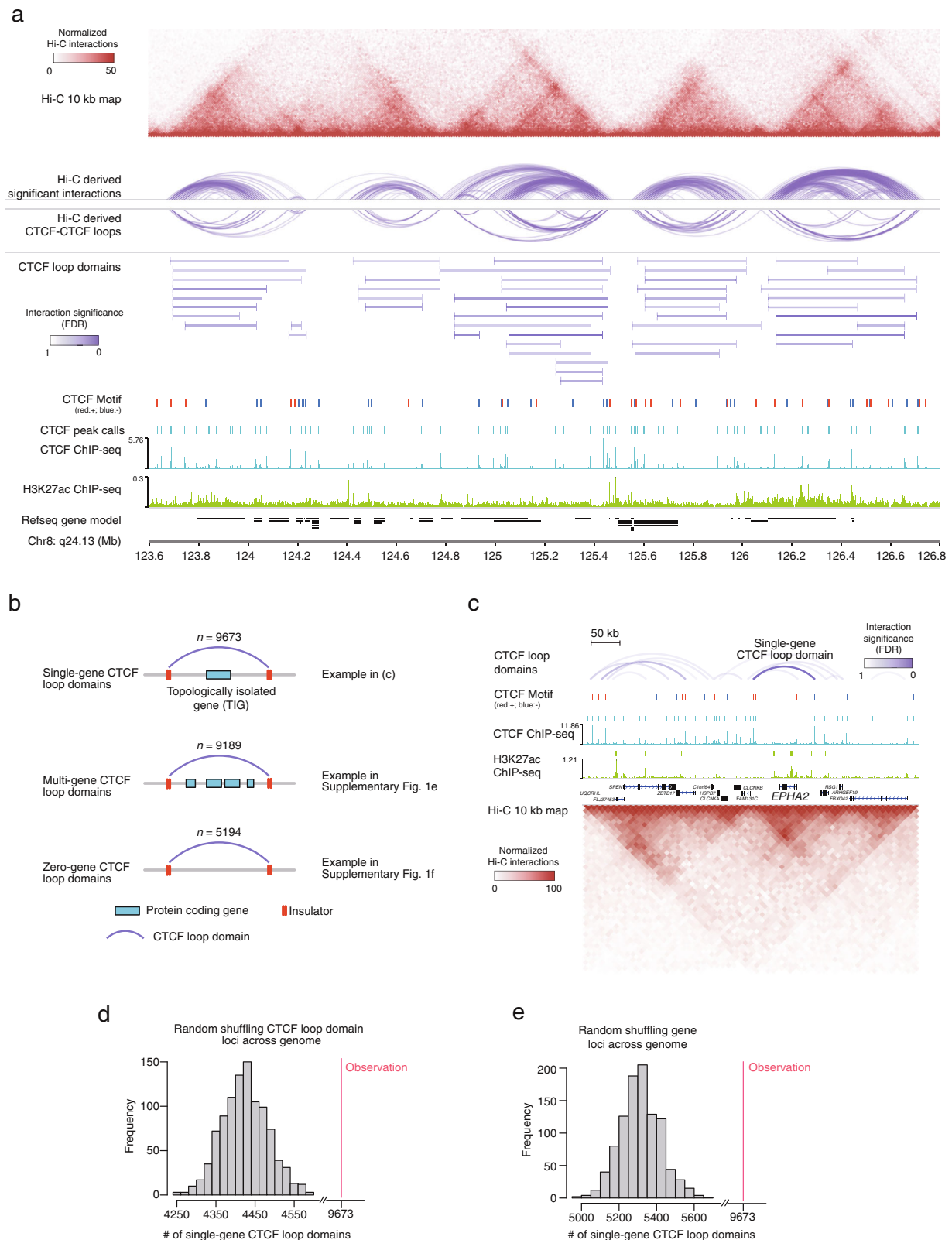
We hypothesized that key developmental regulators, especially factors directing early differentiation, might be shielded through insulated structures from nearby genes to facilitate local regulation, and that disruption of their boundaries might lead to deregulation and consequently cellular defects. To investigate this, we conducted a systematic study, analyzing patterns of insulated domains across the human genome and show that CTCF loop domains<sup>21,22</sup> display an intriguing enrichment that

may facilitate proper regulation of crucial genes. Specifically, we find that early developmental regulators appear preferentially isolated from other genes, with looped boundary CTCF sites that are highly conserved across cell types. To functionally explore their role, we used single guide (sg)RNA Cas9-directed genome perturbation to disrupt the CTCF loop domain boundary at the *SOX17* (an example of a single isolated gene) and *NANOG* (an example of a multi-gene domain) loci. Notably, the boundary of *SOX17*, but not *NANOG*, appears necessary for proper function: disruption led to *SOX17* misexpression and a failure to differentiate into endoderm. Moreover, we found a subset of CTCF loop domains with constitutive boundaries across many cell types, which are also more conserved in sequence across species. Our findings add further support to the contribution of gene insulation through chromosomal folding structures to enable defined expression of developmental regulators, especially during early differentiation.

## Results

**Identification of topologically insulated regions.** To explore the role of topological genome organization in pluripotent stem cells, we performed Hi-C experiments<sup>1</sup> in the human embryonic stem (ES) cell line HUES64 and generated a total of 1.05 billion uniquely mapped paired-end reads (Supplementary Data 1). These data led to the identification of 231,970 high-confidence interactions in the genomic range of 20 kb–2 Mb using Fit-Hi-C<sup>23</sup> (Methods). By mapping these interactions to CTCF consensus motifs, we obtained 37,428 significant CTCF-CTCF loops. Loops close to each other were clustered and merged to limit redundancy, yielding a total of 24,056 CTCF loop domains with a median length of 304 kb (Fig. 1a, Supplementary Fig. 1a and Supplementary Data 2). To further validate our domain calling results, we compared the identified loops to those reported by other Hi-C loop detection methods<sup>24–26</sup> and observed a large degree of agreement between Fit-Hi-C and HiCCUPS results (Supplementary Fig. 1b). Moreover, the CTCF loop domains identified by Fit-Hi-C depicted the highest agreement with the insulated neighborhoods identified by cohesin ChIA-PET data in primed human ES cells<sup>20</sup> (Supplementary Fig. 1c). We then calculated a directionality index, which provides a quantification for the degree of upstream or downstream interaction bias of a genomic region<sup>2</sup>, for all CTCF loop domains and surrounding regions to demonstrate their topological insulation function (Supplementary Fig. 1d). Surprisingly, using this data, we found that many CTCF loop domains contain only a single protein-coding gene (~40%,  $n = 9,673$ ) (Fig. 1b, 1c, and Supplementary Fig. 1e, 1f), which is a significant overrepresentation compared to what is expected (permutation test  $p < 0.001$  when randomly shuffling either domains or genes, Fig. 1d, e). These CTCF loop domains are termed single-gene domains, and the genes contained are referred to as topologically isolated genes (TIGs). The remaining CTCF loop domains embed either multiple genes (38%,  $n = 9,189$ ; Supplementary Fig. 1e) or no genes (22%,  $n = 5,194$ ; Supplementary Fig. 1f), respectively.

**CTCF loop domain boundaries are largely preserved during ES cell differentiation.** We next sought to explore the stability of CTCF loop domain boundaries using an ES cell differentiation model and generated Hi-C data from ES-cell-derived endoderm (dEN), ectoderm (dEC), and mesendoderm (dMS). We found that these boundaries are well preserved during ES cell differentiation as exemplified at the *SOX17*, *SMAD1*, *SOX2*, and *NANOG* loci by visualizing the heatmaps and arc plots of Hi-C interactions (Supplementary Fig. 2a); however, this approach is not suitable for analyzing large numbers of boundaries at the



same time. We therefore used boundary-anchored virtual 4 C plots to visualize contact interactions between all pairs of boundaries in a sample. This approach uses two heatmaps, with the left heatmap representing the Hi-C interactions from the surrounding genomic regions of the left boundaries to the right boundaries (similar to setting the left boundaries as viewpoints in

4 C data) and the right heatmap representing the Hi-C interactions from the surrounding genomic regions of the right boundaries to the left boundaries (Fig. 2a). For cases in which there are physical interactions between the two boundaries in the sample, the plot exhibits a high intensity in the center of both heatmaps but not the surrounding regions. By investigating the

**Fig. 1 Topologically insulated regions in HUES64 ES cells.** **a** Normalized Hi-C interaction map, high-confidence interactions (arc), CTCF-CTCF loops (arc), and CTCF loop domains (line) displayed on top of ChIP-seq profiles of CTCF and H3K27ac at chromosome 8q24.13 region. CTCF peaks are denoted by bars above the ChIP-seq profiles of CTCF. CTCF consensus motifs are denoted by red (forward orientation) and blue (reverse orientation) bars above the CTCF peaks. Normalized Hi-C interactions are shown as a heatmap with each pixel representing a 25 kb genomic region. The interaction significance (FDR) was calculated from Hi-C data. **b** Illustration of single-gene CTCF loop domains, multi-gene domains, and zero-gene domains. The numbers of domains in each group within HUES64 are displayed on top of each plot. **c** Display of a single-gene CTCF loop domain and the topologically isolated gene (TIG) at the *EPHA2* locus. CTCF consensus motifs, ChIP-seq profiles of CTCF and H3K27ac, and normalized Hi-C interaction maps are displayed. CTCF peaks and enhancers are denoted by bars above the ChIP-seq profiles of CTCF and H3K27ac. CTCF consensus motifs are denoted by red (forward orientation) and blue (reverse orientation) bars above the CTCF peaks. Normalized Hi-C interactions are shown as a heatmap with each pixel representing a 10 kb genomic region. CTCF loop domains are displayed on the top and the interaction significance (FDR) was calculated from Hi-C data. **d** The distribution of the number of single-gene domains in the human genome by randomly shuffling the domain loci across the genome. The red line indicates the observed number of single-gene domains in the genome. **e** The distribution of the number of single-gene domains in the human genome by randomly shuffling the gene loci across the genome. The red line indicates the observed number of single-gene domains in the genome.

profiles of both *P*-values (Fig. 2a) and normalized Hi-C contacts (Fig. 2b), we found that the boundaries of CTCF loop domains are largely preserved throughout the process of ES cell differentiation. We then aggregated these heatmaps for each sample by column to generate a profile plot representing the average and standard deviation of the signal across all boundaries. This analysis depicts a clear peak in the boundary centers of both single-gene and multi-gene domains (Fig. 2c and Supplementary Fig. 2b), and demonstrates that single-gene domain boundaries are more preserved than multi-gene domain boundaries (Fig. 2d).

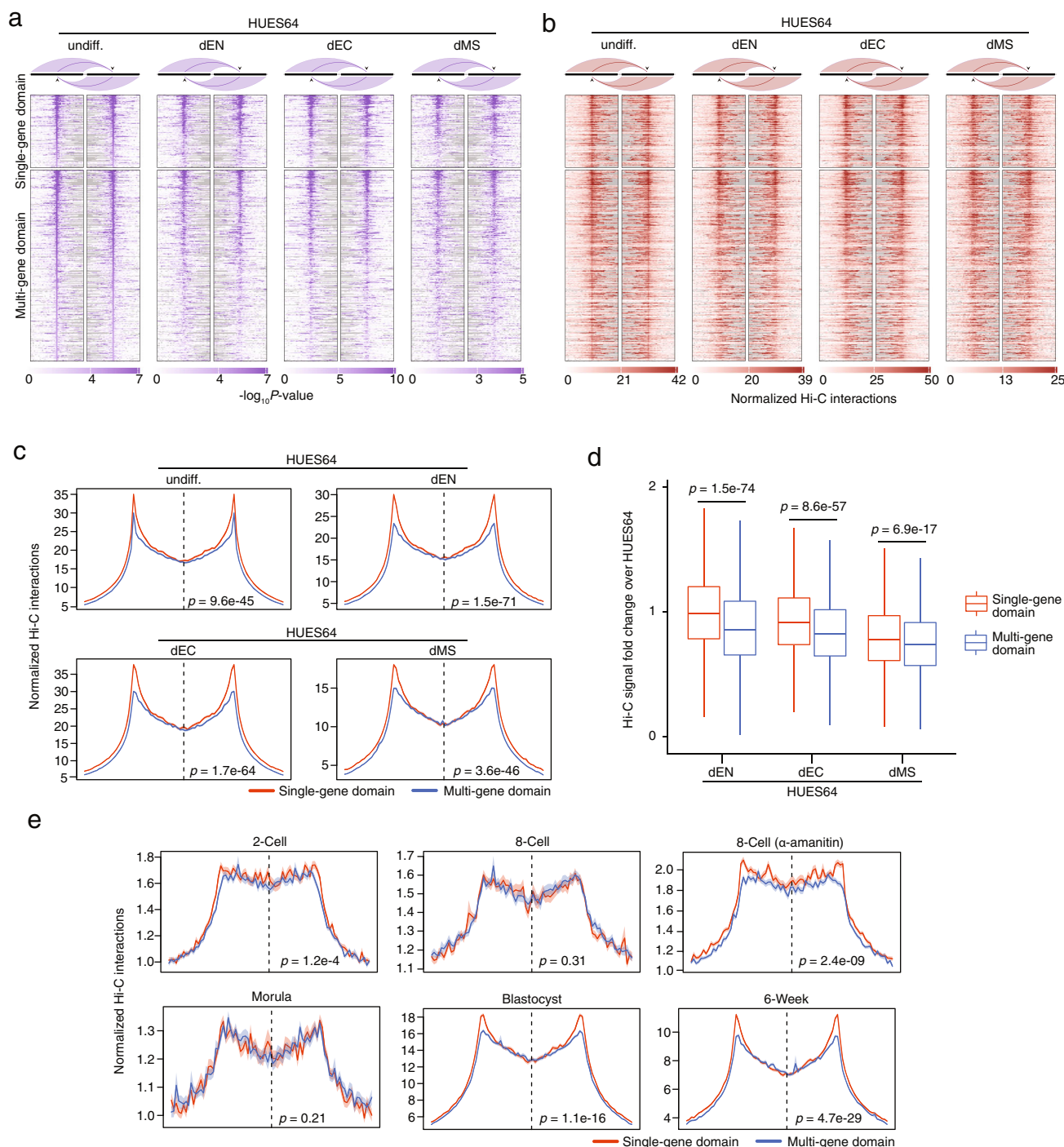
To interrogate when these boundaries are formed during embryonic development, we analyzed recently published Hi-C data from early human embryos, including samples of 2-cell, 8-cell, morula, and blastocyst stages, as well as a 6-week time point<sup>27</sup>. We found that CTCF loop domain boundaries are gradually established starting from the 8-cell stage, at the same time or shortly after zygotic genome activation (ZGA)<sup>28</sup> during which CTCF gene expression is also strongly induced<sup>27</sup>, and are stable after the blastocyst stage (Fig. 2e). At the blastocyst stage and 6-week time point, the single-gene domain boundaries demonstrated more pronounced Hi-C interactions than the multi-gene domain boundaries, as observed in ES cells and their derivatives, independent of genomic distances (Supplementary Fig. 2c). However, this difference was not observed in the early stages (8-cell and morula), implying that this boundary divergence arises between the morula and blastocyst stages and coincides with the initiation of lineage specification<sup>29</sup>. ZGA inhibition by  $\alpha$ -amanitin (an RNA Pol II inhibitor) treatment was shown to also repress CTCF expression in the 8-cell stage<sup>27</sup>. Interestingly,  $\alpha$ -amanitin treatment of 8-cell embryos had less influence on single-gene domain boundaries than on multi-gene domain boundaries (Fig. 2e). Once formed, it appears that single-gene domain boundaries are maintained independently of ZGA and of CTCF expression, and might thus be more stable and robust across diverse cellular processes.

**Developmental regulators are insulated by conserved CTCF boundaries.** The preservation of single-gene domain boundaries in ES cell differentiation may imply some functional importance. Indeed, further analysis demonstrated that TIGs are enriched for diverse developmental processes in the Gene Ontology (GO) database (Fig. 3a). Next, we defined developmental regulators as transcriptional factors under the GO term “developmental process” and performed enrichment analysis of the different CTCF loop domain groups, based on the number of genes they insulate. Interestingly, we found that developmental regulators are enriched for in single-gene CTCF loop domains, but less so in other domain groups with multiple genes (Fig. 3b). This association

motivated us to query whether the insulation function of these boundaries is important.

First, we analyzed the extent of conservation of these boundaries across different mammals and found that boundary CTCF motifs of single-gene CTCF loop domains are highly conserved across placental mammals (Fig. 3c and Supplementary Fig. 3a). This may suggest that these motifs are functionally important elements that undergo natural selection. In contrast, the boundary CTCF motifs of multi- and zero-gene CTCF loop domains have a sequentially decreasing conservation score, while CTCF motifs outside of any boundaries (nonboundary CTCF motifs) are generally not conserved (Fig. 3c and Supplementary Fig. 3a). We also observed that boundary CTCF sites of developmental regulator domains are more conserved than those of other genes (Fig. 3d). These analyses further support the notion that the boundary CTCF sites of single-gene domains, especially those insulating developmental regulators, may be functionally important. In addition, we found that boundaries of single-gene domains are more strongly interacting based on Hi-C data (Supplementary Fig. 3b) and are enriched for stronger CTCF-binding sites than other boundaries, with a similar average signal intensity for single- and multi-gene domain boundaries (Supplementary Fig. 3c-e). Finally, single-gene domains were found to contain more *cis*-regulatory elements per gene, such as enhancers and long noncoding RNAs, than other domains (Supplementary Fig. 3f-g). Taken together, these results demonstrate that TIGs are enriched for developmental regulators and that their conserved boundaries are of potential regulatory importance.

**Single-gene domain boundary perturbation leads to dysregulation.** Next, we sought to investigate the effect of single-gene domain boundaries disruption on the gene they insulate throughout ES cell differentiation. We refined the list of developmental regulators curated from HUES64 cells to specifically include early developmental regulators (eDR), which displayed a stronger enrichment in both CTCF loop domains and TIGs than other developmental regulators (Supplementary Fig. 4a). The majority of these regulators are located within CTCF loop domains (89%, 33/37) (Fig. 4a), 25 of which are TIGs (Supplementary Fig. 4b-d), 8 are in multi-gene domains (Supplementary Fig. 4e), and 4 are located in CTCF loop domain-free regions (Supplementary Fig. 4f). To enable functional characterization of a representative TIG, we chose the *SOX17* locus as it is isolated by strong boundaries (boundary interaction strength adjusted *P*-value =  $3.5 \times 10^{-9}$  based on Hi-C data) and encodes a member of the SOX (SRY-related HMG-box) family of transcription factors that is specifically induced in early endoderm differentiation by distal enhancers<sup>30</sup> (Fig. 4a, b and Supplementary Fig. 2a).



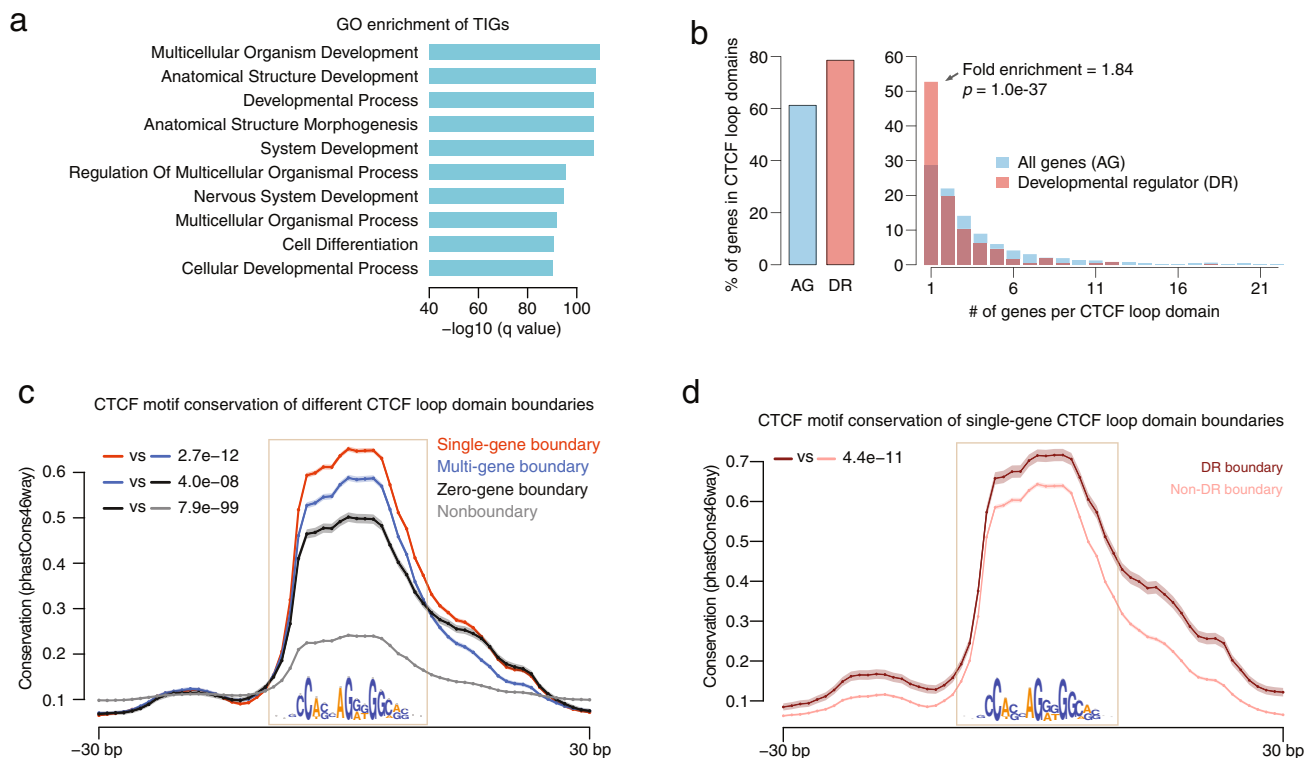
We designed two sgRNAs flanking the 5' centromeric boundary of the *SOX17* CTCF loop domain (Boundary 2), which is about 300 kb away from the locus (Fig. 4b and Supplementary Fig. 5a), and derived three independent homozygous *SOX17* $\Delta 5'$ CTCF clones in the female iPSC line ZIP13K2<sup>31</sup>. The switch to an iPSC line has various practical benefits, such as sharing material and data across labs, and Hi-C data for human ESCs and iPSCs are very similar<sup>32</sup>. We further confirmed that the CTCF occupancy and Hi-C boundary strength at the *SOX17* locus are also similar between human ESCs and iPSCs (Supplementary Fig. 6a).

We then confirmed the deletion of a 5 kb region that includes two CTCF peaks (Fig. 4c, Supplementary Fig. 5a, b) and the

corresponding boundary interaction loss (Boundary 2) along with an increased interaction specifically in definitive endoderm at the upstream boundary (Boundary 1) in one of the boundary deletion cell lines (*SOX17* $\Delta 5'$ CTCF#8.2) (Fig. 4d, e, Supplementary Fig. 5a, b and Supplementary Data 3). We also identified a significant reduction of intraloop domain contacts (*SOX17* loop domain, *SOX17* upstream loop domain) in *SOX17* $\Delta 5'$ CTCF#8.2 iPSCs as well as a reduction of endoderm-specific enhancer contacts between the *SOX17* promoter and its most distal regulatory element DRE (A) (Fig. 4d, e and Supplementary Fig. 6b, c). In concordance with highest intergroup contact correlations (Supplementary Fig. 6d) and without any further evidence of ectopic enhancer adoption or alternative enhancer



**Fig. 2 CTCF loop domain boundaries in ES cell differentiation.** **a** Boundary-anchored virtual 4 C heatmap of the domain boundaries in HUES64 and its derivatives. The locations of domain boundaries were identified in HUES64 Hi-C data. The  $-\log_{10}$   $P$ -value (before adjusting for multiple comparisons) obtained from Fit-Hi-C software are shown. Each row represents the domain of one gene. The strongest domain (i.e., that with the lowest Hi-C interaction  $P$ -value between boundaries) per gene is shown if there are multiple domains for that gene. **b** Boundary-anchored virtual 4 C heatmap of the domain boundaries identified from HUES64 Hi-C data plotted by using the Hi-C data of HUES64 and its derivatives as the underlying contact maps. The normalized Hi-C interactions are shown. The ordering is the same as in **a**. **c** Boundary-anchored virtual 4 C average profile of the domain boundaries in HUES64 and its derivatives. The locations of domain boundaries were identified in HUES64 Hi-C data. The normalized Hi-C interactions are shown. The dotted line separates the left and right boundary regions, which represent the regions in the left and right heatmap in **b**. Average signals across all boundaries are shown with the shaded area indicating the standard error. Two-sided Wilcoxon test was used to determine significance level of boundary-to-boundary interactions between the two groups. Data are presented as mean values  $\pm$  SE. **d** Hi-C signal fold-change of boundary-to-boundary interactions in HUES64 derivatives over HUES64 cells. Two-sided Wilcoxon test  $p$ -value is shown. Hi-C signals were normalized by library size in individual samples prior to the analysis. The box indicates the interquartile range (IQR), the line inside the box shows the median, and whiskers show the locations of either  $1.5 \times$  IQR above the third quartile or  $1.5 \times$  IQR below the first quartile,  $n = 3,310$  boundary-to-boundary interactions for single-gene domain,  $n = 8,729$  boundary-to-boundary interactions for multi-gene domain. **e** Boundary-anchored virtual 4 C average profile of the domain boundaries at the 2-cell, 8-cell, 8-cell treated with  $\alpha$ -amanitin, morula, blastocyst stages, and 6-week embryos. The locations of domain boundaries were identified in HUES64 Hi-C data. CTCF expression is inhibited under  $\alpha$ -amanitin treatment at the 8-cell stage. The normalized Hi-C interactions are shown. The dotted line separates the left and right boundary regions. Average signals across all boundaries are shown with the shaded area indicating the standard error. Two-sided Wilcoxon test was used to determine the significance level of boundary-to-boundary interactions between the two groups. Data are presented as mean values  $\pm$  SE.

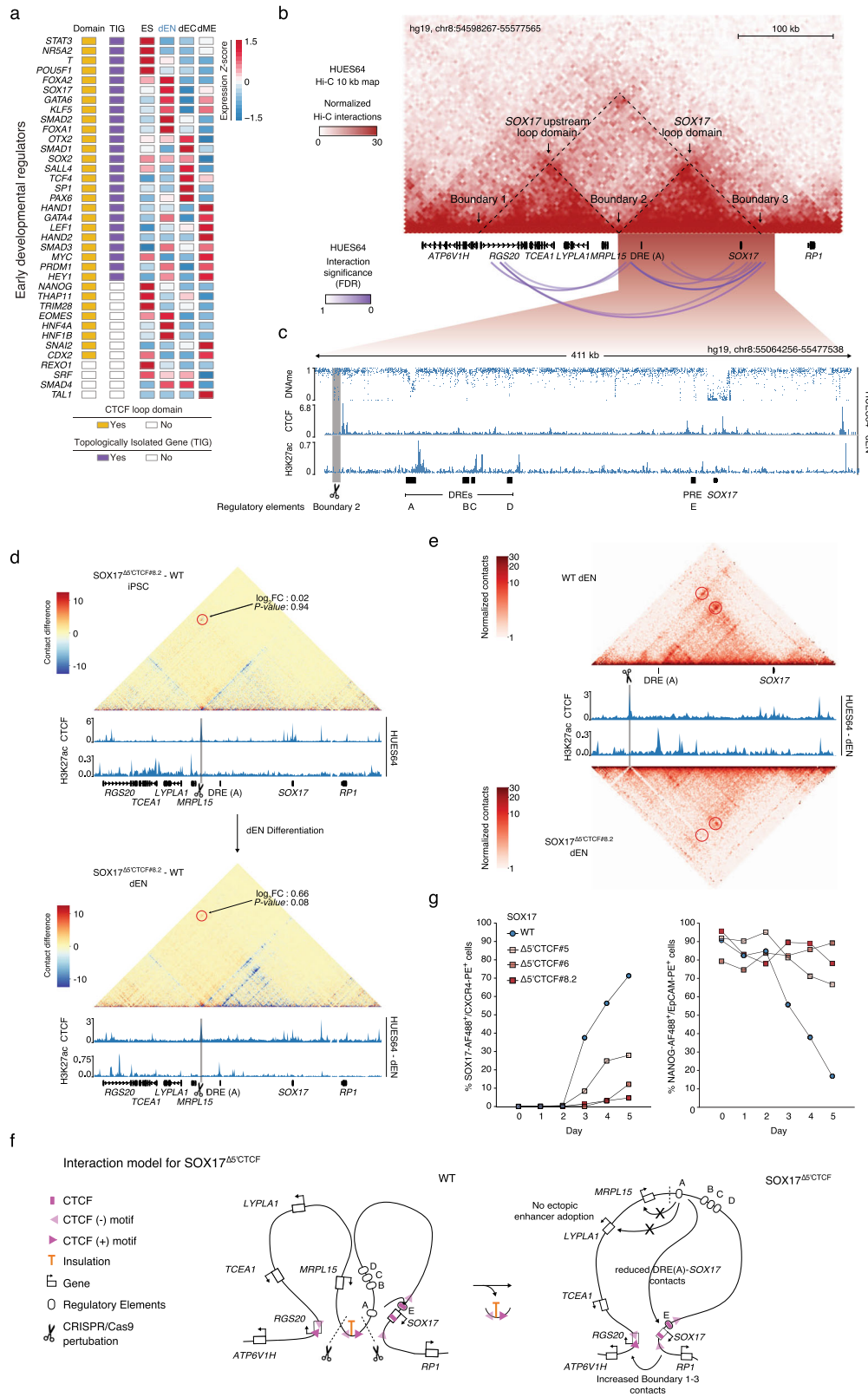


**Fig. 3 Topological insulation of developmental regulators.** **a** Gene Ontology (GO) enrichment of TIGs in HUES64 shows developmental processes as top terms. **b** Enrichment of developmental regulators in single-gene CTCF loop domains. The left panel represents the percentage of all genes (AG) or developmental regulators (DR) located within domains. The right panel represents the percentage of AG or DR located within domains containing an increasing number of protein-coding genes.  $P$ -values calculated by two-sided Fisher's exact test. **c** Evolutionary conservation of consensus CTCF motifs at boundaries of single-gene, multi-gene, and zero-gene CTCF loop domains and nonboundaries. Nonboundary CTCF motifs represent the motifs that are outside of any domain boundaries. The motif region is shown in the box and the motif sequence is displayed. The average conservation score across placental mammals across all boundary regions is shown. The shaded area indicates the standard error. Two-sided Wilcoxon test  $p$ -value tested in the motif regions is shown. Data are presented as mean values  $\pm$  SE. **d** Evolutionary conservation of consensus CTCF motifs at DR boundaries and non-DR boundaries. See more descriptions in **c**. Data are presented as mean values  $\pm$  SE.

looping due to boundary perturbation (Fig. 4d, e), we concluded that there was a decreased frequency of enhancer-gene contacts during definitive endoderm formation between the *SOX17* promoter and its tissue-specific enhancer DRE (A)) in the *SOX17* $\Delta^{5}$ <sup>CTCF#8.2</sup> cell line (Fig. 4f).

*SOX17* is known to be a key early endoderm transcription factor<sup>33,34</sup> and is frequently used to identify embryonic

endodermal tissues, e.g., primitive, visceral, and definitive endoderm<sup>30</sup>. Together with the transmembrane C-X-C chemokine receptor 4 (CXCR4), *SOX17* is used to specifically confirm definitive endoderm cell identity<sup>30,34</sup>. Notably, when we used directed differentiation conditions for generating definitive endoderm, we found a strong reduction in *SOX17*<sup>+</sup>/*CXCR4*<sup>+</sup> cell populations in all *SOX17* $\Delta^{5}$ <sup>CTCF</sup> isogenic clones (on average



4.68–27.95%) compared to wild-type cells (71.3%) (Fig. 4g and Supplementary Fig. 5c, d). To assess whether the mutant cells have already exited pluripotency and lost their epithelial character due to epithelial-to-mesenchymal-transition (EMT)<sup>30,35</sup>, we utilized the pluripotency transcription factor NANOG in combination

with the transmembrane glycoprotein Epithelial Cell Adhesion Molecule (Ep-CAM). We found reduced NANOG<sup>+</sup>/Ep-CAM<sup>+</sup> cell populations only in wild-type cells (16.9%) while population numbers remained comparably high over time in SOX17<sup>Δ5CTCF</sup> (on average 66.7–89.25%, Fig. 4g and

**Fig. 4 Single-gene versus multi-gene domain boundary perturbation highlights TIG-dependent gene regulation.** **a** Heatmap of early developmental regulators displays information on CTCF loop domains, TIGs, and expression in the embryonic stem cell differentiation process. The RPKM (Reads Per Kilobase per Million mapped reads) value of gene expression in embryonic stem (ES) cells, definitive endoderm (dEN), ectoderm (dEC), and mesoderm (dME) were row Z-scored. **b** Multi-layered display of the *SOX17* locus as a representative TIG at chr8:54598267-55577565. HUES64 CTCF loop domains are displayed as arcs below a normalized Hi-C interaction map. **c** Multi-layered display of HUES64 derived dEN WGBS, CTCF, and H3K27ac ChIP-seq profiles. Putative *SOX17* distal regulatory elements (DRE) and proximal regulatory elements (PRE) are highlighted in black bars and given capital letters. The deleted centromeric *SOX17* boundary (Boundary 2) is highlighted in grey marked by a scissor. **d** Capture Hi-C subtraction maps in iPSCs (upper panel) and dEN cells (lower panel) at the *SOX17* locus. The relative contact difference between the two samples (*SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup>wild-type) in either iPSCs or dEN cells are shown on top of HUES64 or HUES64 derived dEN CTCF and H3K27ac ChIP-seq profiles. Boundary 1 + 3 contact quantifications are highlighted in red circles. The deleted centromeric *SOX17* boundary (Boundary 2) is highlighted in grey marked by a scissor. *SOX17* DRE (A) and gene bodies are highlighted in black bars. **e** Capture Hi-C maps in dEN wild-type (upper panel) and *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> (lower panel) at the *SOX17* loop domain. The normalized capture Hi-C contact maps are overlaid with HUES64 derived dEN CTCF and H3K27ac ChIP-seq profiles. Relative contact differences between Boundary 2+3 or between the *SOX17* promoter and DRE (A) are highlighted in red circles. The deleted centromeric *SOX17* boundary (Boundary 2) is highlighted in grey and marked by a scissor. Putative *SOX17* DRE (A) and *SOX17* gene body are shown as black bars. **f** Simplified 2D-model of the *SOX17* boundary 2 perturbation in wild-type or *SOX17*<sup>Δ5<sup>CTCF</sup></sup> cells. Genes are depicted as white rectangles, regulatory elements as white ellipses. Crucial boundary related CTCF-ChIP-seq peaks are shown in pink; available motif-orientations are highlighted as arrows. Insulation is shown in orange. Dashed lines and scissors indicate the predicted Cas9 cut sites at boundary 2. **g** Fluorescence activated cell sorting (FACS) time-course data of wild-type and *SOX17*<sup>Δ5<sup>CTCF</sup></sup> iPSC during directed differentiation towards definitive endoderm. *SOX17* and *CXCR4* (CD184) are depicted as percentage *SOX17*<sup>+</sup>/*CXCR4*<sup>+</sup> in bulk cell populations. Corresponding *NANOG* and *Ep-CAM* (CD326) are depicted as percentage *NANOG*<sup>+</sup>/*Ep-CAM*<sup>+</sup> in bulk cell populations ( $n = 2$  biologically replicates). Data are presented as mean values.

Supplementary Fig. 5c, d), suggesting a boundary-dependent deregulation of *SOX17* gene control and a failure to properly exit pluripotency.

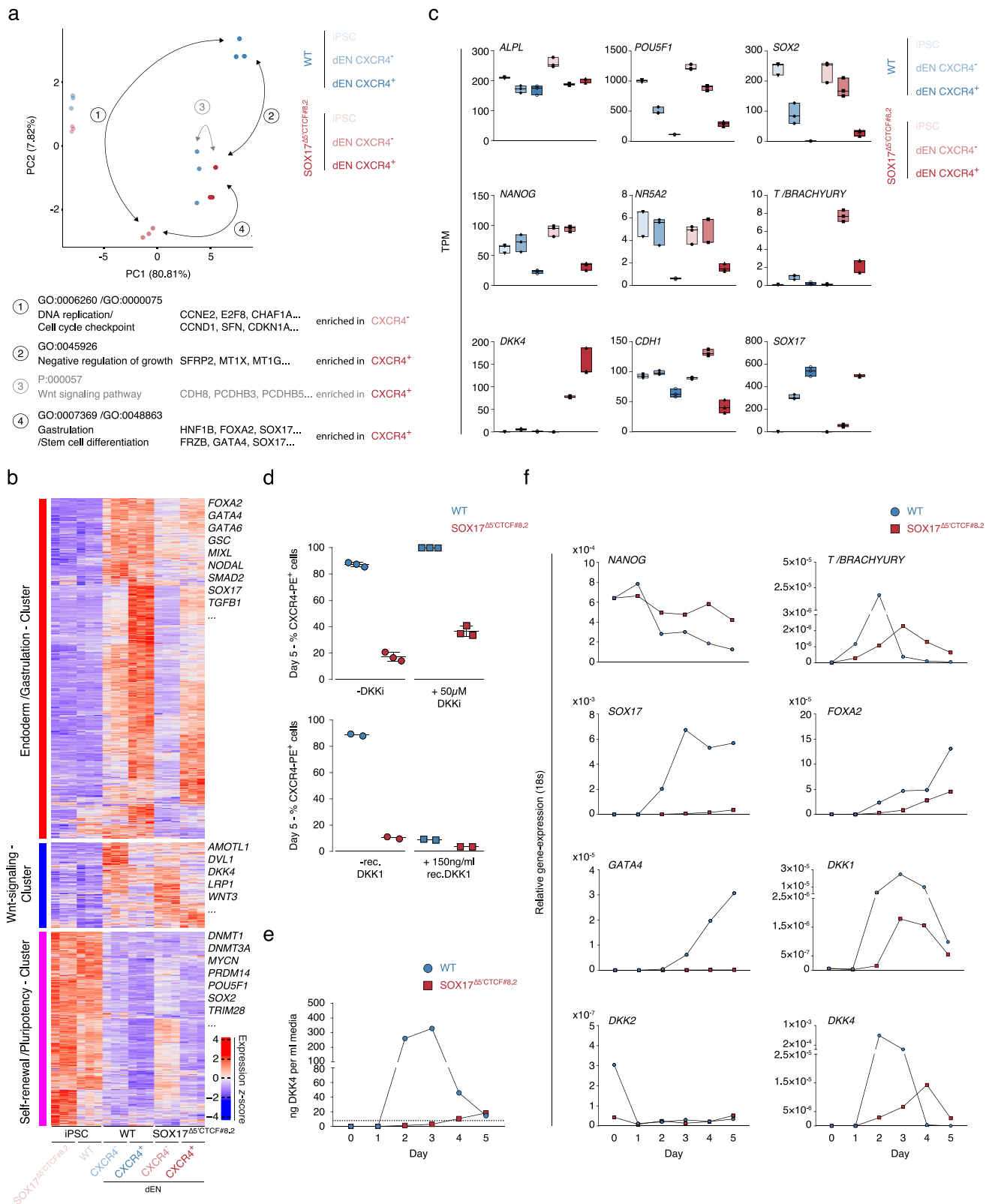
To compare the effect with a multi-gene domain boundary deletion, we chose the *NANOG* locus as it is also isolated by strong boundaries (FDR=5.38e-10) in Hi-C data and encodes a highly expressed pluripotency TF in human iPSCs (Fig. 4a and Supplementary Fig. 5e)<sup>36</sup>. Two sgRNAs flanking the 3' centromeric boundary of the *NANOG* CTCF loop domain were designed, about 20 kb away from the *NANOG* locus (Supplementary Fig. 5e, f). We derived one homozygous and two heterozygous *NANOG*<sup>Δ3<sup>CTCF</sup></sup> isogenic clones in the female iPSC line ZIP13K2<sup>31</sup>, in which deletions of a 2 kb region including one CTCF motif were further confirmed (Supplementary Fig. 5f, g). Interestingly, we did not observe deregulation of relative *NANOG* protein and mRNA levels in *NANOG*<sup>Δ3<sup>CTCF</sup>#21</sup> compared to wild-type cells (Supplementary Fig. 5h, i). mRNA expression levels of all other genes localized within the *NANOG* CTCF loop domain were also not found to be altered in *NANOG*<sup>Δ3<sup>CTCF</sup>#21</sup> compared to wild-type cells (Supplementary Fig. 5i). Thus, we identified an important role for single-gene domain boundary-dependent regulation of genes as indicated by the *SOX17* locus and the fusion of both *SOX17* CTCF loop domains highlighted by a strongly disrupted differentiation outcome.

***SOX17* boundary perturbation leads to endoderm differentiation failure.** Next, we sought to confirm the absence of CTCF in *SOX17*<sup>Δ5<sup>CTCF</sup></sup> cells and performed CTCF-ChIP qRT-PCR on *SOX17* Boundary 2 and control regions (Fig. 4b, c). We observed a genotype-specific loss of CTCF occupancy within and spanning *SOX17* Boundary 2 in *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> compared to wild-type iPSCs (Supplementary Figs. 5a and 7a). Due to the loss of this insulator in *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> cells, we next aimed to assure the absence of a potential adoption of the *SOX17* DRE (A) by upstream genes in an endoderm-specific context (Fig. 4b, c). Therefore, we isolated different *CXCR4* subfractions for RNA-seq followed by differential gene-expression analysis; using this approach, we confirmed normal regulation of *SOX17* CTCF loop domain-associated genes in dEN, except for *SOX17* (Supplementary Fig. 7b). Expression of *SOX17* was exclusively observed in *CXCR4*<sup>+</sup> *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> cells, which comprised only a minor fraction of the population (on average 4.68–27.95%) (Fig. 4g and

Supplementary Fig. 7b). This data suggest a boundary-dependent deregulation of *SOX17* that is not associated with ectopic DRE-adoption by *SOX17* CTCF upstream loop domain-related genes.

To gain more insights into the transcriptomes of differentiated populations, we performed principle component analysis (PCA) of the 100 most variable genes across all samples (Fig. 5a and Supplementary Fig. 7c-g). Interestingly, *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> *CXCR4*<sup>+</sup> and wild-type *CXCR4*<sup>-</sup> cell populations closely clustered together on an endodermal differentiation trajectory roughly between undifferentiated and *CXCR4*<sup>+</sup> wild-type populations (Fig. 5a). Since *CXCR4*<sup>+</sup> wild-type and *CXCR4*<sup>-</sup> *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> populations comprised the respective majority, we analyzed differentially expressed genes ( $n = 1,506$ ) using GSEA for biological processes ( $\log_2FC > 2$ ,  $q$ -value  $< 0.05$ ) and found genes enriched for DNA replication and cell cycle checkpoint in *CXCR4*<sup>-</sup> *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> cells (Supplementary Fig. 7d). As described previously, determination of endodermal cell fate propensity is closely connected to the cell cycle<sup>37</sup>, suggesting that *CXCR4*<sup>-</sup> *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> cells are transcriptionally delayed and about to enter definitive endoderm. When comparing wild-type and *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> *CXCR4*<sup>+</sup> populations (437 genes), we found genes associated with negative regulation of growth to be enriched in *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> cells (Supplementary Fig. 7e). Interestingly, when comparing the more differentiated *CXCR4*<sup>+</sup> *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> population with their *CXCR4*<sup>-</sup> counterpart (635 differentially expressed genes), we found genes associated with gastrulation and stem cell differentiation (Supplementary Fig. 7f) enriched in *CXCR4*<sup>+</sup> cells, which again points towards a developmental delay of *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> *CXCR4*<sup>-</sup> cells and a compromised ability to generate proper definitive endoderm.

***DKK4* deregulation implicates a WNT signaling defect in *SOX17*<sup>Δ5<sup>CTCF</sup></sup> cells.** Surprisingly, pathway GSEA led to the identification of enriched WNT signaling in *SOX17*<sup>Δ5<sup>CTCF</sup>#8.2</sup> *CXCR4*<sup>+</sup> compared to wild-type *CXCR4*<sup>-</sup> cells (Fig. 5a and Supplementary Fig. 7g). To further perform paired gene enrichment analysis in an unbiased way, we performed expression z-score clustering of the most variable genes ( $n = 4,151$ ) throughout all groups (Fig. 5b). We defined three clusters: the endoderm/gastrulation cluster (2,282 genes), WNT signaling cluster (569 genes), and self-renewal/pluripotency cluster (1,300



genes). Genes associated with the endoderm/gastrulation cluster were most highly expressed in CXCR4<sup>+</sup> wild-type cells while self-renewal/pluripotency cluster genes were found to be most highly expressed in both iPSC populations (Fig. 5b). Interestingly, WNT signaling-associated genes were most highly expressed in both CXCR4<sup>-/+</sup> SOX17<sup>Δ5</sup>CTCF#8.2 but also CXCR4<sup>-</sup> wild-type

populations, confirming our previous GSEA enrichment analysis (Fig. 5a, b). One of these marker genes, *DKK4*, was found to be exclusively upregulated in SOX17<sup>Δ5</sup>CTCF#8.2 cell populations (Fig. 5c). *DKK4* has been shown to antagonize canonical WNT signaling by the inhibition of LRP5/6 interaction with WNT, forming a ternary complex with the transmembrane protein

**Fig. 5 SOX17 boundary perturbation leads to endoderm differentiation failure.** **a** Principal component analysis of RNA-seq data, depicting sample clusters by the use of the 100 most variable genes. The first two principal components (PCs) are displayed. Arrows and numbers indicate group comparisons. GSEA of differentially expressed genes between compared groups are indicated below; significantly enriched biological processes are depicted in black, pathways in gray. **b** TPM Z-score row normalized clustering of the most variable genes ( $n = 4,151$ ). **c** TPM values shown for a subset of genes ( $n = 3$  biological replicates). The box indicates the interquartile range (IQR), the line inside the box shows the median. **d** Wnt stimulation/antagonization utilizing fluorescence activated cell sorting (FACS) data of wild-type and SOX17 $\Delta^5$ CTCF iPSC at day 5 definitive endoderm. CXCR4 (CD184) is depicted as percentage CXCR4 $^+$  in bulk cell populations. The upper panel shows DKK2/3/4 inhibition and controls (treatment for 5 consecutive days) ( $n = 3$  biological replicates). Data are presented as mean values  $\pm$  SD. The lower panel depicts human recombinant DKK1 treatment and controls (for 5 consecutive days) ( $n = 2$  biological replicates). Data are presented as mean values. **e** DKK4 Enzyme-linked Immunosorbent Assay (ELISA), a quantitative measure of human DKK4 in cell culture supernatants over time ( $n = 3$  biological replicates (averaged) over 2 experiments). Data are presented as mean values. **f** qRT-PCR of bulk populations from a subset of genes related to Wnt signaling, mesendoderm, endoderm, and pluripotency over 5 days endoderm differentiation. Expression values are depicted as relative gene-expression ( $2^{-\Delta Ct(GOI-18s)}$ ) ( $n = 2$  biological replicates). Data are presented as mean values.

KREMEN that promotes internalization of LRP5/6<sup>38</sup>. Hence, expression of DKK4 may lead to insufficient canonical WNT signaling required for proper endodermal differentiation<sup>39</sup>.

To explore the relevance of DKK4, we utilized a chemical inhibitor compound (9-Carboxy-3-(dimethyliminio)-6,7-dihydroxy-10-methyl-3H-phenoxazin-10-ium iodide), which led to partially rescued CXCR4 $^+$  bulk population levels (on average 36.4%) (Fig. 5d). As a control experiment, we considered WNT inhibition instead by utilizing recombinant DKK1, which led to notably reduced CXCR4 $^+$  bulk populations in wild-type (on average 8.95%) but not SOX17 $\Delta^5$ CTCF#8.2 cells (Fig. 5d). To test DKK4 levels released into the culture medium, we performed Enzyme-Linked Immunosorbent Assay (ELISA) over 5 days of dEN differentiation. We found a striking reduction of DKK4 levels in SOX17 $\Delta^5$ CTCF#8.2 culture supernatants compared to wild-type over time (Fig. 5e): DKK4 release was found to be slowly increasing and delayed over time, indicating no impact of WNT inhibition during differentiation, but rather being a consequence of deregulated SOX17 gene control. The importance of WNT signaling, its role in endoderm and the functional relation between SOX17 and WNT signaling was demonstrated in studies utilizing *Xenopus* gastrulation<sup>40,41</sup>. Hence, we suggest a functional lack of SOX17 $\Delta^5$ CTCF#8.2 cells to respond properly to WNT signaling, most likely due to the boundary perturbation-dependent deregulation of SOX17.

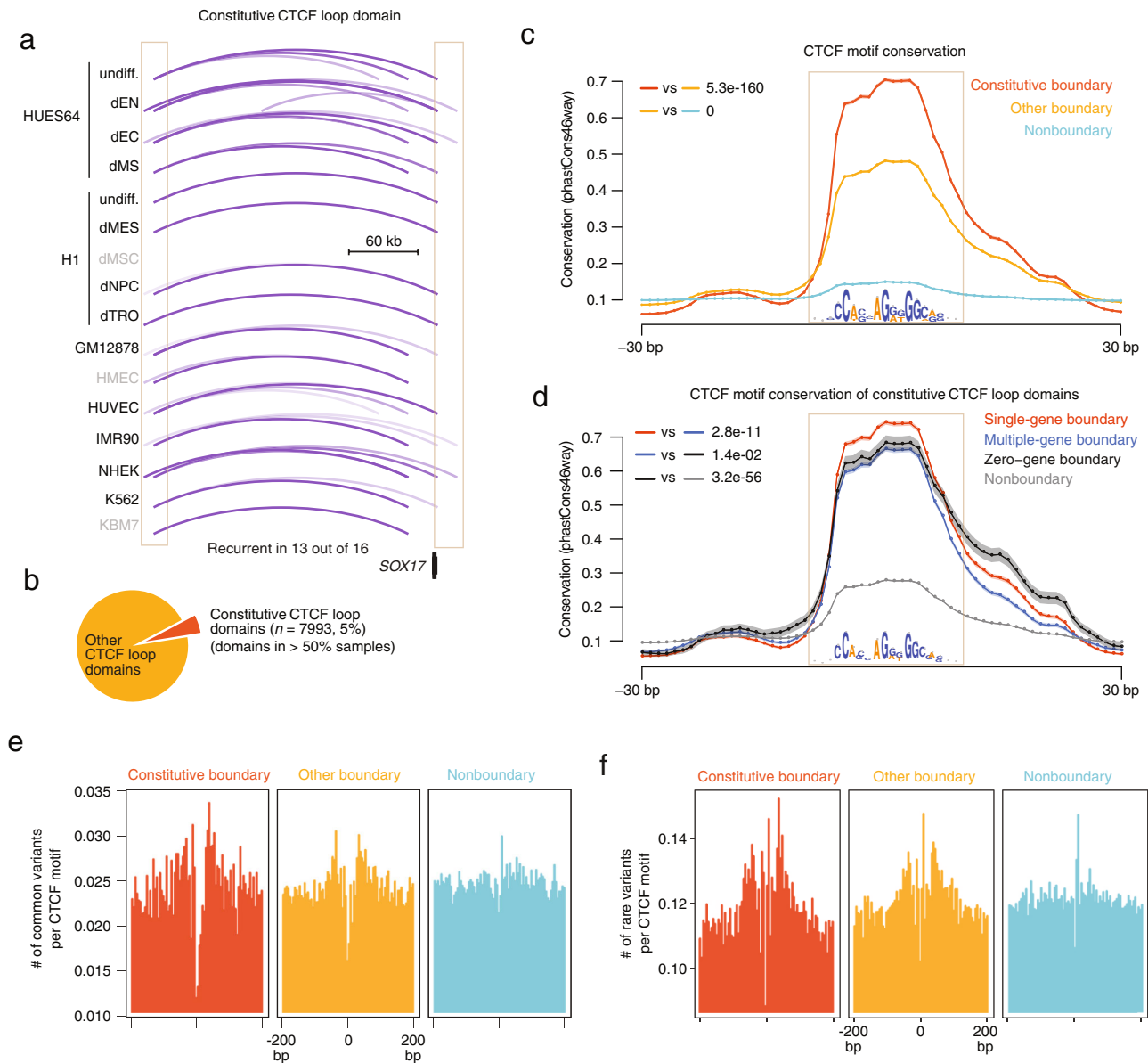
In contrast, SOX17 $\Delta^5$ CTCF CXCR4 $^-$  cell populations highly express mesendodermal markers, such as *T/BRACHYURY* and *NR5A2*, prematurely accompanied by high levels of pluripotent markers such as *NANOG*, *SOX2*, and *POU5F1* but not *ALPL1*. We also obtained elevated levels of the key epithelial marker *CDH1* as well as very low levels of *SOX17* to be expressed (Fig. 5c and Supplementary Fig. 7c). To test whether SOX17 $\Delta^5$ CTCF#8.2 cells may exert a rather delayed endoderm differentiation program, we performed qRT-PCR analysis on time-course differentiated bulk populations. We found strong expression reduction of the endodermal markers *SOX17*, *FOXA2*, and *GATA4* in concordance with stable *NANOG* expression over time, as confirmed by previous FACS data (Figs. 5f and 4g). We also found that the mesendodermal regulator *T/BRACHYURY* and WNT antagonists *DKK1/DKK4* showed reduced expression early in differentiation at day 1–3 and a clear expression onset delay of around 1–2 days, resulting in elevated expression levels at day 5 (Fig. 5f).

To investigate whether the observed phenotype is reversible by restoring the endoderm-required SOX17 expression levels, we made use of a destabilized ectopic SOX17-TagBFP expression system, which we randomly integrated into the SOX17 $\Delta^5$ CTCF#8 genetic background using the Piggy-BAC transposase (Supplementary Fig. 8a). Hygromycin-selected and TagBFP $^-$ -sorted SOX17 $\Delta^5$ CTCF#8 cells were further cultured in bulk and named SOX17 $^{DDSOX17}$ . Endoderm differentiating SOX17 $^{DDSOX17}$  cells were either treated

(SOX17 $^{DDSOX17+}$ ) or not treated (SOX17 $^{DDSOX17-}$ ) with a small molecule<sup>42</sup> named Shield-1 from day 2 onwards to either reverse the constitutive ectopic SOX17-TagBFP degradation or not, (Supplementary Fig. 8b). To first explore leakiness of our system, we performed a western blot assay, which revealed some minimal ectopic SOX17-TagBFP degradation in SOX17 $^{DDSOX17-}$  undifferentiated iPSCs but also terminally differentiated dEN cells (Supplementary Fig. 8b). Interestingly, we found not only elevated levels of ectopic but also endogenous SOX17 protein in day 5 differentiated SOX17 $^{DDSOX17+}$  cells compared to SOX17 $^{DDSOX17-}$  cells, indicating a coupled activation of the endogenous SOX17 locus (Supplementary Fig. 8b). From day 2 of endoderm differentiation onwards, we observed CXCR4 $^+$  fractions to be restored to almost wild-type levels in SOX17 $^{DDSOX17+}$  (Supplementary Fig. 8d). Surprisingly, we even found increased CXCR4 $^+$  populations in SOX17 $^{DDSOX17-}$  cells compared to the original knockout SOX17 $\Delta^5$ CTCF#8 cells, again indicating leakiness of our expression system (Supplementary Fig. 8d). Although SOX17 $^{DDSOX17-}$  cells were found to be leaky for ectopic SOX17-TagBFP degradation, we observed a retained NANOG-expressing fraction of cells compared to SOX17 $^{DDSOX17+}$  by immunofluorescence staining (Supplementary Fig. 8e).

Finally, to explore if the extent of transcriptional rescue in SOX17 $^{DDSOX17+}$  and CXCR4 $^{+/-}$  cells would resemble wild-type gene expression, we performed PCA of the 100 most variable genes across wild-type, SOX17 $\Delta^5$ CTCF#8 and SOX17 $^{DDSOX17+}$  cells and found both populations including the undifferentiated iPSCs closely clustering with their wild-type matching cell populations (Supplementary Fig. 8f). In sum, our results suggest that SOX17 $\Delta^5$ CTCF cells can still exit pluripotency, but are delayed and trapped in a mesendoderm-like state due to WNT signaling nonresponsiveness via deregulated SOX17, leading to the eventual endoderm differentiation failure reversible by ectopic SOX17 expression.

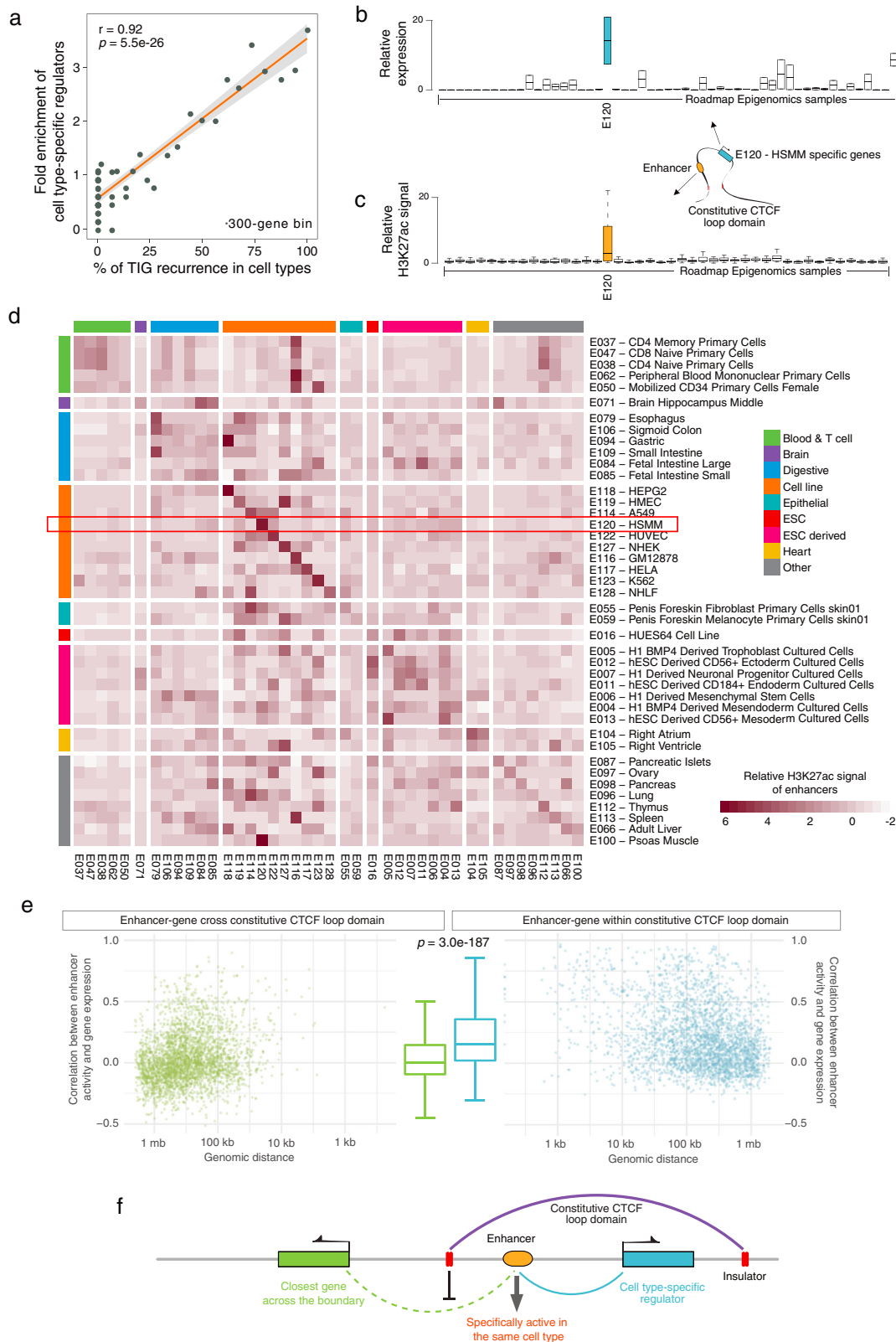
**Constitutive CTCF loop domains and their essential functional roles.** After exploring the functional relevance of the boundaries of SOX17 and NANOG, we aimed to more generally explore the function of boundaries genome-wide. As previously reported, CTCF-CTCF loops are preserved in different cell types<sup>5,10,12,43</sup>, and constitutive loops are functionally important in tumors<sup>10</sup>. Therefore, we analyzed a collection of high-resolution Hi-C data obtained from 16 different cell lines and defined constitutive CTCF loop domains as those that were detected in at least 50% of all samples (Fig. 6a, b and Supplementary Data 4–5). We observed significant enrichment of early developmental regulators (18/37; fold enrichment = 5.24;  $p$ -value =  $7.67e-10$ ) among single-gene constitutive CTCF loop domains, with SOX17, SMAD2, GATA4, STAT3, LEF1, FOXA2, and KLF5 as top representatives (Supplementary Fig. 9a). As expected, we found that the CTCF boundaries of constitutive CTCF loop domains are more conserved than those identified from individual cell types (Fig. 6c);



**Fig. 6 A functionally essential role of constitutive CTCF loop domains. a** Display of a constitutive domain at the SOX17 locus that is conserved across 13 out of 16 human cell types. Arcs represent the domains in different cell types. The constitutive boundaries are shown via the box. **b** The pie plot shows the proportion of constitutive domains and domains. Constitutive domains are present in >50% of samples, while domains are present in  $\leq$ 50% of samples. **c** Evolutionary conservation of consensus CTCF motifs at constitutive domain boundaries, domain boundaries, and nonboundaries. Nonboundary CTCF motifs are those motifs not in any domain boundaries. The motif region is shown in the box and the motif sequence is displayed. The average conservation score across placental mammals across all boundary regions is shown. The shaded area indicates the standard error. Two-sided Wilcoxon test  $P$ -values tested in the motif regions are shown. Data are presented as mean values  $\pm$  SE. **d** Evolutionary conservation of consensus CTCF motifs at single-gene, multi-gene, and zero-gene boundaries of constitutive domains and nonboundaries. Nonboundary CTCF motifs are the motifs not in any constitutive domain boundaries. See more descriptions in **c**. Data are presented as mean values  $\pm$  SE. **e** The number of common variants per consensus CTCF motif site at constitutive domain boundaries, domain boundaries and nonboundaries. Common variants are genetic variants with an allele frequency larger than 1% in the 1000 genome project phase 3 data. **f** The number of rare variants per consensus CTCF motifs at constitutive domain boundaries, domain boundaries, and nonboundaries. Rare variants are genetic variants with allele frequency less than 1% in 1000 genome phase 3 data.

furthermore, TIG boundaries within-constitutive domains were more conserved than others (Fig. 6d). We therefore hypothesized that, if domain boundaries are essential for controlling the expression of the developmental regulators they contain, the disruption of such elements should be negatively selected for in individuals. To test this hypothesis, we analyzed the 1000 Genome Project phase 3 data, which contains 84.4 million variants identified from data on 2,504 individuals from 26 human populations, to interrogate frequent variants in CTCF loop domain

boundaries. We observed a depletion of common variants (allele frequency >1% in the population) in domain boundaries relative to flanking regions, especially for constitutive CTCF loop domain boundaries (Fig. 6e); nonboundary CTCF sites did not demonstrate such a depletion. Rare variants (allele frequency <1% in the population) showed a similar but less pronounced trend (Fig. 6f). These findings indicate that boundaries of constitutive domains, if altered, are subject to purifying selection, thus supporting their essential role in preventing deleterious phenotypes.



**Co-activation within and insulation across constitutive CTCF loop domains.** To further explore how these boundaries of constitutive domains influence the expression of genes they contain, we utilized the Roadmap Epigenomics Project dataset, which includes measurements of both gene expression (RNA-seq) and enhancer activity (H3K27ac ChIP-seq) in the same set of

cells. First, we defined cell-type-specific regulators by selecting transcription factors that demonstrate cell-type-specific expression patterns across 54 tissues and cell lines (Methods). We observed that TIG recurrence in cell lines was positively correlated with enrichment of cell-type-specific regulators (Fig. 7a,  $r = 0.92$ ,  $P$ -value =  $5.5e-26$ ). For consistency, we identified

**Fig. 7 Co-activation within and insulation across constitutive CTCF loop domains.** **a** Fold enrichment of cell-type-specific regulators against the recurrence of TIGs in multiple cell types. Each point represents 300 genes. Linear regression line with 95% confidence interval in light gray is shown. *P*-values calculated by two-sided Pearson correlation test. **b** Relative expression of cell-type-specific regulators in E120 across all cell types. Relative expression is the expression within a cell type normalized by the mean expression across all cell types. The box indicates the interquartile range (IQR), the line inside the box shows the median, and whiskers show the locations of either  $1.5 \times$  IQR above the third quartile or  $1.5 \times$  IQR below the first quartile,  $n = 2$  genes. **c** Relative H3K27ac abundance of enhancers located within the same constitutive domain of the cell-type-specific regulators in E120 across all cell types. Relative H3K27ac abundance is the H3K27ac signal over mean H3K27ac signal across all cell types, which represents enhancer activity. The box indicates the interquartile range (IQR), the line inside the box shows the median, and whiskers show the locations of either  $1.5 \times$  IQR above the third quartile or  $1.5 \times$  IQR below the first quartile,  $n = 28$  enhancers. **d** Mean H3K27ac abundance of enhancers located within the same constitutive domain of the cell-type-specific regulators. Each row depicts the mean of the relative H3K27ac abundance of enhancers located within the same constitutive domain of the cell-type-specific regulators in the corresponding cell type across all cell types. The red box indicates the mean value of the boxplot in **c**. Note that some tissues and cell lines are functionally related, which may drive the enrichment off the diagonal, such as hematopoietic cell-type-specific enhancers being also enriched in GM12878 and thymus. Enhancers specific to cell lines are stronger and uniquely enriched in a specific cell type, such as HEPG2 enhancers. Some enhancers show tissue-type-specific properties instead of cell type specificity, such as enhancers of ESCs and derivatives, hematopoietic cells, and heart cells. **e** Correlation between enhancer activity and within-constitutive domain-neighbor gene expression (left plot in green) and correlation between enhancer activity and cross-constitutive domain-neighbor gene expression (right plot in blue). The boxplot shows the whole distributions of the data shown on the left and right. The box indicates the IQR, the line inside the box shows the median and whiskers show the locations of  $1.5 \times$  IQR above the third quartile and  $1.5 \times$  IQR below the first quartile,  $n = 3,611$  enhancers. Note that only cell-type-specific regulators (TIGs) and their constitutive domains are included in this analysis. **f** Model of co-activation of enhancer and cell-type-specific regulators within the same single-gene constitutive domain. The schematic also depicts the insulation function of constitutive domain boundaries.

constitutive TIGs and investigated their overlap with cell-type-specific regulators (Supplementary Fig. 9b). We found that those genes and enhancers localized within the same constitutive domain were co-activated in either a cell-specific or a tissue-specific manner (Fig. 7b–d and Supplementary Fig. 10). Further analyses demonstrated that enhancer activity is more correlated with its neighbor gene expression within-constitutive domains than outside of them (Fig. 7e, Wilcoxon test *P*-value < 0.0001). These results support the presumed insulation function of constitutive domains by restricting the enhancer activity to the targeted genomic region, and further demonstrate that topological isolation can add to precise control of local gene expression (Fig. 7f).

## Discussion

Elucidating the relationship between chromosome structure and gene regulatory programs is of broad interest. Increasing evidence has demonstrated that mediator cohesin loops, mostly enabling enhancer-promoter interactions, have substantial effects on gene regulation in diverse systems<sup>7–9,20,43–46</sup>. Similarly, CTCF cohesin loops, mostly functioning as insulators, have been proposed to constrain enhancer-promoter interactions for proper gene-expression patterns<sup>5,10,12,14,47</sup>. Here, we describe an aspect of genome organization that may facilitate the precise temporal and spatial control of key developmental regulators. Our model is supported by functional data showing that disruption of a CTCF loop domain boundary can strongly impact the lineage commitment of pluripotent cells. Integrative and systematic analyses further highlight sequence conservation, germline variant and boundary constraint profiles, which extend our functional case study and demonstrate the importance of CTCF loop domain boundaries on a genome-wide scale. This understanding of topological isolation and the precise control it may exert on developmental processes suggests a potential mechanism of co-evolution between transcriptional control and chromosome structure formation. Our work supports a model suggesting that gene duplication and its subsequent organization in its own single-gene domain may be a frequent way to evolve and acquire new gene functions without disrupting neighboring genes and their regulatory elements.

A previous study demonstrated that knockouts of CTCF loop domain boundaries leads to altered expression of nearby genes in

mouse ES cells, providing evidence that the maintenance of topological boundaries is important for the proper expression of these genes<sup>5</sup>. However, there is limited knowledge how boundary perturbation-induced gene-expression changes influence ES cell differentiation in humans. Our functional study on the *SOX17* locus demonstrated that disruption of CTCF loop domain boundaries strongly impacts cell lineage commitment of human iPS cells. In addition, enhancer adoption/hijacking after boundary perturbation has been widely observed and extensively studied in both development and tumorigenesis<sup>10,14,29,47</sup>. Here, we shown that the boundary knockout at the *SOX17* locus did not induce enhancer adoption by other genes but causes loss of proper enhancer regulation of its endogenous targets. Our results suggest a dual function of topological insulation—the boundary interaction not only constrains enhancer activity within the domain, but also facilitates enhancer-promoter interaction by bringing them into physical proximity (Fig. 4g). This observation may also imply the existence of diverse mechanisms of topological insulation<sup>17</sup>, which need to be further dissected.

We demonstrate more pronounced boundary conservation of CTCF loop domains containing a single gene than those containing multiple genes, suggesting a possibly more critical role of boundaries of single-gene than those of multi-gene domains. This observation is in-line with findings of our boundary disruption experiment at the *NANOG* locus, which is a multi-gene CTCF loop domain, in ES cells, where we did not observe any phenotypic change and no genes within the locus had significantly altered expression levels. Since *NANOG* is highly expressed and required for the maintenance of pluripotency in ES and iPS cells<sup>36,48</sup>, these results suggest that gene-expression regulation in this context is independent of boundary disruption in pluripotent cells. These results provide a preliminary understanding of the difference between single-gene and multi-gene CTCF loop domains, with the limitation of the currently still relatively small number of reported examples<sup>9,10,14,15,17,47,49</sup>. Validation of our findings with additional CTCF loop domain boundary functions in different cell states is needed to arrive at a better understanding of how 3D topological structures control the gene-expression program in multi-cellular processes. One of the barriers preventing such a functional study is the lack of a uniform phenotype; for instance, cell viability may not be a good indicator for all cellular processes<sup>50</sup>, and in some cases, boundary disruptions of a CTCF



loop domain have not led to immediate gene-expression changes<sup>51</sup> and clear phenotypes<sup>49</sup>. Thus, it might be important to consider the specific context and the exact point in time when topological insulation may exert its control on gene expression<sup>52</sup>.

Our study has demonstrated that 90% of early developmental regulators, many of which are essential, are localized in CTCF loop domains; this high representation led us to hypothesize that these regulators need to be shielded from interference by neighboring regulatory elements or need their own elements within an isolated topological domain to be accessible. By contrast, there might be more flexibility for regulators acting in somatic cells, as deregulation of such factors may cause a developmental disorder or tumorigenesis but might not be immediately lethal. In both cases, the boundary alterations via DNA mutations or structural variations might be able to be used as diagnostic markers or therapeutic targets across multiple disease types. For instance, a previous study demonstrated that forced chromatin looping by tethering enhancers to repressed  $\gamma$ -globin genes reactivated their expression by overriding the endogenous topological structures<sup>53</sup>. Another study developed a light-activated system to conduct endogenous gene-expression control via dynamic induction of artificial chromosome loops<sup>54</sup>. These recent technologies provide promising therapeutic approaches to treat diseases caused by 3D topological alterations, potentially leading to the emergence of 3D therapeutics.

## Methods

**Parameters.** Default parameters were used, if not otherwise specified, for all software and pipelines utilized in this study.

**Hi-C sequencing.** Hi-C libraries were prepared following the protocol described in Rao et al.<sup>7</sup>. Briefly, one million cells were crosslinked with 1% formaldehyde for 10 min at room temperature and then quenched with 0.2 M glycine solution. Cells were lysed and nuclei permeabilized with 0.5% sodium dodecyl sulphate for 10 min at 62 °C. Chromatin was digested with 100 U of MboI restriction enzyme (New England Biolabs). Ends of the restriction fragments were blunted and labeled with a biotinylated nucleotide and then ligated. Nuclei were pelleted, proteins were digested with proteinase K and crosslinks were reversed by heating at 68 °C overnight. DNA was sheared in a Covaris focused ultrasonicator to average fragment length of 400 bp. Size-selected DNA was enriched for biotinylated ligation products through binding to T1 streptavidin beads (Thermo Fisher). Libraries were prepared for Illumina sequencing by performing the end repair, A-tailing and adapter ligation steps with DNA attached to the beads. Hi-C libraries were amplified directly off the beads and purified for subsequent Illumina sequencing with 100 paired-ends.

**Identification of CTCF loop domains from Hi-C data.** Raw Hi-C reads were mapped to the hg19 version of the human genome and preprocessed using the Hi-C-Pro pipeline<sup>55</sup> (version 2.11.0-beta) to obtain uniquely mapped deduplicated interactions. These interactions were then aggregated into 10 kb genomic bins and normalized using the calCB algorithm in HiCapp<sup>56</sup> (v1.0.0). The high-confidence (i.e., significant,  $q < 0.01$ ) interactions in the genomic range of 30 kb–2 Mb were identified using the Fit-Hi-C python package<sup>23</sup> (v1.0.1). By mapping the anchors of high-confidence interactions to CTCF sites (the union of CTCF motifs and CTCF-ChIP-seq peaks in the corresponding sample), we obtained CTCF-CTCF loops. We observed that some samples with low sequence depth had very few identified loops, because small counts led to a low power for interactions to pass the significance cutoff. Therefore, we applied a hard cutoff to obtain the top 10,000 CTCF-CTCF loops in these samples based on previous evidence regarding the number of these loops per cell type<sup>5,7,10</sup>. Subsequently, loops close to each other were clustered and merged to reduce redundancy (see below). These merged loops generated the final set of CTCF loop domains. Note that summits of merged loops were identified based on the Hi-C interaction significance and were used instead of the merged loops themselves to increase the resolution of anchor points. The same procedure was performed by using three other Hi-C loop detection methods with the recommended parameter settings by the original references: HiCCUPS<sup>25</sup> (-m 500 -r 10000 -f.1), SIP<sup>24</sup> (-res 10000 -fdr 0.05), and Homer<sup>26</sup> (-res 10000 -window 50000). The CTCF loop domains were compared across different methods, and were compared to insulated neighborhoods identified by cohesin ChIA-PET data in primed human ES cells<sup>20</sup>.

**Identification of topologically isolated genes (TIGs) from Hi-C data.** Protein-coding genes (PCGs) were extracted from the RefSeq annotation of the hg19

version of the human genome. The transcription start sites (TSSs) of PCGs were compared to the localization of CTCF loop domains to decide how many PCGs are located within each CTCF loop domain. CTCF loop domains containing one PCG were named single-gene CTCF loop domains; CTCF loop domains containing more than one PCG are referred to as multiple-gene CTCF loop domains; and CTCF loop domains containing no PCGs were named zero-gene CTCF loop domains. The PCGs in the single-gene CTCF loop domains are referred to as Topologically Isolated Genes (TIGs). The use of the phrase “isolated” is meant to represent that a gene is localized by itself in a CTCF loop domain. To classify genes into these categories, we used the TSS instead of the whole gene body because the promoter region represents the key transcriptional regulator of a gene; it is the promoter that requires tight regulation by insulation of chromosomal regions in order to prevent mis-regulation by nearby elements<sup>5</sup>.

**Boundary-anchored virtual 4 C visualization of Hi-C data.** To visualize the boundary interactions of many CTCF loop domains in a Hi-C data, we used the boundary-anchored virtual 4 C plot. It's a simple way to visualize the interactions between one boundary to the surrounding regions of the other boundary. More specifically, the left heatmap shows the Hi-C interactions between the surrounding genomic regions of the left boundaries and the right boundaries; The right heatmap shows the Hi-C interactions between the surrounding genomic regions of the right boundaries and the left boundaries. Any Hi-C matrix-like scores derived from Hi-C data can be shown by using such plot, such as the normalized Hi-C interactions and the Fit-Hi-C  $p$ -values. Then, the heatmaps can be aggregated by the columns of the left and right heatmaps to generate a shaded line plot with the line represents the average signal across columns and the shaded area represents the standard deviation signal across columns. The shaded line plot could be used to visualize the difference between multiple groups of interactions as well as calculate statistics. The left and right heatmaps are plotted into single shaded line plot with a dotted vertical line to separate them.

**Identification of constitutive CTCF loop domains and TIGs from multiple Hi-C datasets.** We used the CTCF loop domains identified from 16 Hi-C datasets to obtain the union CTCF loop domains across cell types (Supplementary Data 4). One key step in CTCF loop domain calling is to use CTCF-binding sites to filter high-confidence interactions identified from Hi-C data, because Hi-C interactions may contain other types of chromosomal structures such as enhancer-gene loops that do not belong to CTCF cohesin loops. In practice, if particular CTCF peaks fail to be detected in the CTCF ChIP-seq data of one or several samples, even if Hi-C data were to show a high-confidence interaction loop at that genomic position, we would miss the CTCF-CTCF loop in these samples. To avoid this scenario, we used the same consensus CTCF-binding sites for each sample instead of the binding sites obtained from individual ChIP-seq data to identify CTCF loop domains in the constitutive CTCF loop domain analysis. The constitutive CTCF loop domains were then defined as those CTCF loop domains that were identified in at least 50% of all cell types (see above), and the genes located within single-gene constitutive CTCF loop domains are referred to as constitutive TIGs (cTIGs).

**Clustering and merging of redundant loops.** We designed a two-step iterative clustering algorithm to cluster and merge paired-end loops within a certain genomic range cutoff; here we used a 1 kb region of boundary overlaps. In the preclustering step, we ranked all loops by their chromosome position and subsequently divided them into two groups based on whether they had even or odd ranks. We then used the pairToPair command in bedtools<sup>57</sup> (v2.25.0) to investigate the overlaps of boundaries between any paired loops from the two sets. The loops in one set that overlapped with any loops in the other set were merged to form new loops with union boundary regions. The loops in one set having no overlaps with any loops in the other set were retained. The merged and retained loops were used as the input for the next iteration. We iteratively applied this process  $N = 50$  times to obtain preclustered loops. In the complete-clustering step, we used the same strategy as in the previous step, except for searching for the overlaps between the preclustered loops and other loops in the same set, instead of dividing them into two loop sets. Self-pairs were excluded from the analysis. In this step, the iterations were continued until the algorithm converges and no paired-end loops can be merged anymore. This two-step procedure was able to cluster and merge a large number of redundant loops in any given genomic range cutoff in a short time period.

**Identification of CTCF motifs and their conservation across species.** CTCF motif loci and orientations in the hg19 version of the human genome were identified using FIMO<sup>58</sup> (v4.11.1). For this analysis, we used the consensus CTCF motif MA0139.1 from the JASPAR CORE 2016 vertebrates database<sup>59</sup>. Motif conservation information was obtained from the UCSC “phastCons46wayPlacental” track.

**Evolutionary analysis of human CTCF loop domain boundaries.** The CTCF motif coordinates of human CTCF loop domain boundaries were lifted over to 45 vertebrate genomes with parameter: -minMatch = 0.9. The motifs successfully lifted over were called present, otherwise absent, in the corresponding genome. The

percent of present motifs in different CTCF loop domain groups across species were studied.

**Identification of consensus CTCF-binding sites.** The CTCF ChIP-seq peaks in 142 different cell lines and tissues (Supplementary Data 5), which were identified using the same settings and contained at least 10,000 peaks, were downloaded from the Cistrome database<sup>60</sup>. The CTCF peaks ( $p < 1e-9$ , peak significance over input) detected in more than 30% of all unique cell types were defined as consensus CTCF-binding sites. The coordinates of ChIP-seq peaks were overlaid with CTCF motifs to obtain orientation information and highest resolution of CTCF-binding sites. Specifically, for the ChIP-seq peaks overlapping with CTCF motif(s), the motif coordinates and orientations were used instead of the peak coordinates. For the ChIP-seq peaks not overlapping with any CTCF motif, the peak coordinates were used and the orientations were set as 'unclear'.

**Clustered and typical CTCF-binding sites.** CTCF ChIP-seq data were analyzed in a similar way as the enhancer analysis of the ROSE pipeline<sup>61</sup>. Specifically, CTCF peaks were merged within a maximal distance of 12.5 kb. The merged peaks were ranked by increasing total ChIP-seq signal, and plotted against the total ChIP-seq signal. This plot showed a clear transition point in the distribution of CTCF occupancy where the total signal began increasing rapidly. The transition point was the  $x$ -axis point for which a line with a slope of 1 was tangent to the curve. We then defined peaks above this point to be clustered CTCF-binding sites, and peaks below that point to be typical CTCF-binding sites. Thus, clustered CTCF-binding sites represent those sites with broad and high CTCF occupancy, while typical CTCF-binding sites represent sites with narrow and low CTCF occupancy.

**Identification of cell-type-specific regulators.** The gene-expression matrix providing transcripts per million (TPM) for 57 samples was downloaded from the Roadmap Epigenomics project<sup>62</sup>. Sample E000, which represents the universal human reference, and three redundant samples (E056, E058 and E061) were removed from the analysis. The gene-expression matrix of the remaining 53 samples (Supplementary Data 5) was then used to identify cell type specifically expressed genes as previously described<sup>63</sup>. We employed the dataset of the Roadmap Epigenomics Project to identify cell-type-specific regulators because it contains diverse cell and tissue types such as stem cells, differentiated cells, primary cells, tissues and immortalized cell lines. The cell-type-specific regulators were defined as transcriptional regulators that were highly induced in certain cell types. Specifically, genes were selected that met the following criteria: (i) entropy less than 0.8; the entropy is calculated as:  $\text{prop} < -x/\text{sum}(x)$ ;  $\text{Entropy} = -\text{sum}(\text{prop} * \log(\text{prop}), \text{na.rm} = \text{T})/\log(\text{length}(\text{prop}))$ , where  $x$  is the TPM vector across samples; and (ii) the maximal TPM across samples is larger than 10; in larger than 0 and less than or equal to 5 samples the gene has at least 7-fold higher expression than the average expression of the gene in all cell types.

#### Identification of enhancers and analysis of their H3K27ac enrichment.

Enhancers were collected from both the Fantom5 database<sup>64</sup> and the Roadmap Epigenomics Project<sup>62</sup>. The enhancers from Fantom5 were directly downloaded from the website ([http://slidebase.binf.ku.dk/human\\_enhancers/](http://slidebase.binf.ku.dk/human_enhancers/)). The enhancers from the Roadmap Epigenomics project were identified from H3K27ac ChIP-seq data of 98 samples. The aligned reads from 98 samples were downloaded as bed files and converted to bam and bigwig files using MACS2<sup>65</sup> (v2.1.0.20150731) and bedtools<sup>57</sup> (v2.25.0). Narrow peaks ( $p < 1e-9$ , peak significance over input) in the same samples called from MACS2 were downloaded and used to identify enhancers via the ROSE pipeline<sup>61</sup> (v0.1). The enhancers from both databases were merged and the union sets were designated as the final list of enhancers. Average H3K27ac signals for enhancers were obtained from the ChIP-seq bigwig files and normalized to signals per 10 million reads for each library. Forty-five of 98 samples for which both H3K27ac ChIP-seq and RNA-seq data was available were used for the analysis (Supplementary Data 5). Enhancers with a maximal signal of less than 5 across all 45 samples were treated as inactive enhancers and removed from the analysis. We found that our results were robust when this cutoff was changed.

**Gene sets and enrichment analyses.** Developmental regulators were genes overlapping with transcription factors and genes under the GO term GO:0032502 - developmental processes. Early developmental regulators were obtained from Tsankov et al.<sup>34</sup>. Gene set enrichment analysis was performed using Fisher's exact test.

**Germline variants from 1000 Genome Project phase 3 data.** We analyzed the 1000 Genome Project phase 3 data, which contains 84.4 million variants identified from data on 2504 individuals from 26 populations, to interrogate frequent variants in CTCF loop domain boundaries. Common variants are genetic variants with an allele frequency larger than 1%; and rare variants are genetic variants with an allele frequency less than 1%.

**Data visualization.** Juicebox<sup>66</sup> was used to generate .hic files of Hi-C data to visualize in the WashU EpiGenome Browser<sup>67</sup> to create the genome track figures.

Other figures were plotted in the R environment (<https://www.r-project.org>) using basic plotting functions and packages of ggplot2<sup>68</sup>, pheatmap (<https://cran.r-project.org/web/packages/pheatmap/index.html>). APA plots were generated by Juicebox<sup>66</sup>.

**Molecular cloning of boundary knockout constructs.** For CRISPR/Cas9 mediated targeting of either *SOX17* or *NANOG* boundary knockout constructs we utilized pSpCas9(BB)-2A-GFP (PX458), which was a gift from Feng Zhang (Addgene plasmid # 48138; <http://n2t.net/addgene:48138>; RRID: Addgene\_48138)<sup>69</sup>. Prior to small guide RNA (sgRNA) cloning, pX458 was initially modified and further renamed into 2X\_pX458. 2X\_pX458 harbors an additional independent U6-promoter followed by a small guide RNA (sgRNA) scaffold expression cassette, which allows the insertion of an additional sgRNA by SapI restriction enzyme cloning. To generate 2X\_pX458, pX458 and the synthesized SapI sgRNA expression cassette (IDT, find sequence below) were digested with KpnI (New England Biolabs, R3142S). Next, the SapI sgRNA expression cassette was ligated into the KpnI linearized pX458 in a 3:1 molarity ratio using T4 DNA-ligase (New England Biolabs, M0202S) according to the manufacturer's instructions followed by transformation and Sanger sequencing to verify successful cloning.

sgRNA-cloning was performed with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, E2621S) according to manufacturer's instructions using BbsI-linearization of 2X\_pX458 for the first sgRNA and SapI linearization of 2X\_pX458 for the second sgRNA as backbone, combined with single stranded oligonucleotides containing the sgRNA sequences as inserts (1:3 molar ratio) (Supplementary Table 1). Bacterial transformation and Sanger sequencing was performed to verify successful cloning. Empty 2X\_pX458 was deposited on addgene.org under ID #172221. The 2X\_pX458 derived *SOX17* and *NANOG* boundary knockout constructs were deposited on addgene.org under ID #172225 and ID #172224 respectively.

**Cell culture and CRISPR/Cas9 targeting.** mTeSR1 (Stemcell Technologies) maintained ZIP13K2 (ref. <sup>31</sup>) human induced pluripotent stem cells were treated with Accutase (Sigma-Aldrich, A6964), supplemented by 10  $\mu\text{M}$  Y-27632 (Tocris, 1254) for 15 min at 37 °C, 5% CO<sub>2</sub> to obtain single cells. To quench and wash the cells, equal volumes of mTeSR1 were added and cells spun down for 5 min at 300  $\times$  g, 21 °C. Cells were further seeded in mTeSR1 containing 10  $\mu\text{M}$  Y-27632 at a density of  $3 \times 10^5$  /cm<sup>2</sup> on Matrigel (Corning) precoated six-well plates (Corning) and cultured 16–24 h at 37 °C, 5% CO<sub>2</sub> before transfection. Transfection was carried out with up to 5  $\mu\text{g}$  of modified P2X458 (including both respective sgRNAs) using Lipofectamine 3000 (Thermo Fischer Scientific) according to the manufacturer's protocol. GFP<sup>+</sup> hiPSCs were FACS-sorted 16–24 h post-transfection on the FACS Aria II (Beckton Dickinson) and seeded in low density ( $5\text{--}10 \times 10^5$ /55 cm<sup>2</sup>) using mTeSR1 supplemented with 10  $\mu\text{M}$  Y-27632 (Tocris, 1254) to derive isogenic clones. Single-cell-derived colonies were picked, and half kept for maintenance respectively used for genotyping with the Phire Animal Tissue Direct PCR Kit (Thermo Fischer Scientific) accordingly. Genotypes were verified by cloning QIAquick Gel Extraction Kit (Qiagen) purified PCR products (Supplementary Table 2) into the pJET1.2 backbone (Thermo Fischer Scientific) and sanger sequencing of PCR single-products was performed with at least 10 $\times$  positively transformed 10-beta *E. coli* (NEB, C3019H) colonies.

**Endoderm differentiation, DKK4 inhibition and DKK1 treatment.** ZIP13K2 cultures were treated with Accutase supplemented by 10  $\mu\text{M}$  Y-27632 to obtain single cells. To quench and wash the cells, equal volumes of mTeSR1 were added and cells spun down for 5 min. at 300  $\times$  g, 21 °C. After resuspension in mTeSR1 supplemented by 10  $\mu\text{M}$  Y-27632, cells were counted and seeded according to the manufacturer's instructions on Matrigel (Corning) precoated culture plates/dishes. Media change using the STEMdiff Trilineage Endoderm Differentiation media was performed on a daily base after washing the cultures with equal volumes of DPBS (Thermo Fischer Scientific, 14190250) according to the manufacturer's instructions. In the case of DKK4 inhibition, differentiation media was supplemented with 50  $\mu\text{M}$  DKK4 inhibitor 9-Carboxy-3-(dimethyliminio)-6,7-dihydroxy-10-methyl-3H-phenoxazin-10-ium iodide (Merck, 317701). In the case of DKK1 treatment, differentiation media was supplemented with 150 ng/ml recombinant human DKK1 (R&D Systems, 5439-DK-010/CF).

**DKK4 enzyme-linked immunosorbent assay (ELISA).** Cell culture media supernatants from undifferentiated or differentiated cells across several timepoints (see *Endoderm differentiation*) of different cell lines were collected, spun at 300  $\times$  g, 5 min at 4 °C. Cell free supernatants were again collected and snap frozen at  $-80$  °C in dry ice. Prior to ELISA, supernatants were thawed on ice and prediluted 1:200 in reagent diluent (R&D Systems, DY995) of the Human Dkk4 DuoSet ELISA KIT (R&D Systems, DY1269). DKK4 ELISA has been carried out according to the manufacturer's instructions. Cell culture media supernatants from undifferentiated or differentiated cells across several timepoints (see *Endoderm differentiation*) of different cell lines were collected, spun at 300  $\times$  g, 5 min at 4 °C. Cell free supernatants were again collected and snap frozen at  $-80$  °C in dry ice. Prior to ELISA, supernatants were thawed on ice and prediluted 1:200 in reagent diluent

(R&D Systems, DY995) of the Human Dkk4 DuoSet ELISA KIT (R&D Systems, DY1269). DKK4 ELISA has been carried out according to the manufacturer's instructions. HRP raw values were measured on the GloMax-Multi Detection System (Promega).

**FACS and Immunofluorescence staining.** Undifferentiated or differentiated ZIP13K2 cultures were treated with Accutase (Sigma-Aldrich, A6964) to obtain single cells. To quench and wash the cells, suspensions were supplemented with FACS buffer containing final 5 mM EDTA (ThermoFischer Scientific, 15575020), 10% fetal bovine serum (FBS) (ThermoFischer Scientific, 26140079) in 1 × DPBS (Thermo Fisher Scientific, 14190250). Further, cells were washed and surface stained in FACS buffer for 30 min at 4 °C using antibody dilutions according to the manufacturer's instructions with slight modifications (Supplementary Table 3). Cells were again washed as described above, fixed and intracellularly stained utilizing the True-Nuclear™ Transcription Factor Buffer Set (Biolegend, 424401) according to manufacturer's instructions. Following subsequent wash steps in permeabilization buffer, we performed flow cytometry data acquisition on the Celesta (Beckton Dickinson, IC-Nr.: 68186, Serial-Nr.: R66034500035). Raw data were analyzed by the use of FlowJo (Beckton Dickinson) v10.7.2.

Undifferentiated or differentiated cell cultures for immunofluorescent (IF) stainings were directly fixed on the culture plates, using 4% PFA solution in DPBS for 15 min at 21 °C. Followed by multiple wash steps with DPBS, cultures were permeabilized in PBT-buffer containing 1% BSA (Sigma-Aldrich, A2153), 10% FBS (ThermoFischer Scientific, 26140079) and 0.3% Triton-X-100 (Sigma-Aldrich, T8787) in DPBS for 30 min at 21 °C. Blocking was further performed in PB buffer (PBT without Triton-X-100) for 30 min at 21 °C. Subsequently, cultures were washed in DPBS and incubated with primary or secondary antibody solutions for at least 2 h at 21 °C (Supplementary Table 4). DNA staining was performed using 0.25 µg/ml DAPI solution (ThermoFischer Scientific, D1306) for 15 min at 21 °C. Microscopy was performed using the Z1 Observer (Zeiss) and fluorescent raw signals were adjusted according to the respective controls using ZEN 2 blue (Zeiss) V2.3. Cell quantification and MFI measurements of the respective channel were performed using Fiji (65, 255 threshold; watershed function; 0.05-0.50 particle size).

**Generation of a polyclonal SOX17-TagBFP cell line and rescue of endogenous SOX17 protein.** PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA rescue construct was generated by Gibson Assembly® (NEB, E2621L) of BstBI /BamHI double-digested PB-CAG-BGHpA (Addgene Plasmid #92161) and EcoRI digested synthetically generated pUC19 DD-3xFLAG-SOX17-GS-TagBFP (Genewiz). PB-CAG-BGHpA was a gift from Xiaohua Shen (Addgene plasmid # 92161; <http://n2t.net/addgene:92161>; RRID:Addgene\_92161)<sup>70</sup>. PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA rescue construct was deposited on addgene.org under ID #172226. Both, PB-CAG-DD-3xFLAG-hSOX17-GS-TagBFP-BGHpA and Super PiggyBac transposase expression vector (SBI, PB210PA-1) were co-transfected into SOX17<sup>Δ5'CTCF</sup>8.2 mTeSR1 (Stemcell Technologies) maintained human induced pluripotent stem cells harboring the SOX17 boundary 2 deletion. Transfection was conducted using equimolar plasmid ratios in combination with Lipofectamine Stem Transfection Reagent (Thermo Fisher Scientific, STEM00003) according to the manufacturer's instructions. Transfected or untransfected cells were treated with mTeSR1 (Stemcell Technologies) containing 250 µg/ml m Hygromycin B (Carl Roth, 1287.1) for 2 weeks. TagBFP-negative surviving cells were FACS-sorted on the FACS Aria Fusion (Beckton Dickinson) and seeded in low density (5–10 × 10<sup>5</sup>/55 cm<sup>2</sup>) using mTeSR1 supplemented with 10 µM Y-27632 (Tocris, 1254) to derive a polygenic/polyclonal SOX17 rescue cell line. To stabilize ectopic SOX17-TagBFP protein, undifferentiated iPSC or day 2 dEN onwards differentiating cells were treated with 1 µM final Shield-1 (Takara, 632189) back to back with untreated controls before sample collection for downstream analysis.

**Western Blot.** Undifferentiated or differentiated ZIP13K2 cultures were treated with Accutase for 15 min, 37 °C, 5% CO<sub>2</sub> to obtain a single suspension. Single-cell suspensions were washed once with ice cold DPBS and spun down at 300 × g, 5 min at 4 °C. Supernatants were removed and cell lysates generated using treatment for 30 min on ice with RIPA buffer (Thermo Fisher Scientific) supplemented with 1 × HALT protease inhibitor (Thermo Fisher Scientific, 87786). Lysates were spun down at 12,000 × g, 10 min at 4 °C and supernatants quantified for protein content using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, 23227) according to the manufacturer's instructions. For western blot, 10 µg total protein extract per sample were boiled in final 1 × Laemmli Buffer (BioRad, 1610747) containing 10% 2-Mercaptoethanol (Sigma-Aldrich) for 10 min at 95 °C, followed by cooling on ice for 5 min. Samples were then loaded on a NuPAGE 4–12%, Bis-Tris, 1.0 mm, Mini Protein Gel (Thermo Fisher Scientific, NP0322BOX) and ran at 200 V for 30 min in 1 × NuPAGE MOPS SDS Running Buffer (Thermo Fisher Scientific, NP0001) containing 1:400 NuPAGE Antioxidant (Thermo Fisher Scientific, NP0005). Protein transfer has been performed utilizing the iBlot 2 Starter Kit, PVDF (Thermo Fisher Scientific, IB21002S) following the manufacturer's instructions for the P0 program. PVDF membranes containing transferred proteins were incubated in blocking buffer (1 × TBS-T (Thermo Fisher Scientific), 5% Blotting-Grade Blocker (BioRad, 1706404)) for 1 h at RT. Incubation with primary antibody dilution (see below) was performed in blocking buffer at 4 °C overnight.

The following day, membranes were washed three times 10 min at RT with 1 × TBS-T and incubated for 2 h at RT in secondary antibody dilution in blocking buffer (Supplementary Table 5). Next, membranes were washed three times for 10 min at RT with 1 × TBS-T and developed using the SuperSignal West Dura Extended Duration Substrate (Thermo Fisher Scientific, 34075) according to the manufacturer's instructions on the BioRad ChemiDoc XRS+ imaging system.

**SureSelect design.** The library of SureSelect enrichment probes were designed over the genomic interval (hg19, chr8:54735936-55657612) using the SureDesign online tool of Agilent. 3299 total probes cover the SOX17 locus and were designed to specifically enrich for regions in proximity of NlaIII sites. The probes covered 35.25% of the interval.

**Capture Hi-C (cHi-C) sequencing and data analysis.** cHi-C libraries were prepared from wild-type or homozygous SOX17<sup>Δ5'CTCF</sup> iPSC or dEN cells. Undifferentiated or day 5 differentiated ZIP13K2<sup>31</sup> cells were grown to a final count of 4–5 million, treated with Accutase (Sigma-Aldrich, A6964), resuspended and washed in DPBS. Cell lysis, NlaIII (NEB R0125) digestion, ligation, and decrosslinking were performed according to the Franke et al. protocol<sup>15</sup>. Adaptors were added to DNA and amplified according to Agilent instructions for Illumina sequencing. The library was hybridized to the custom-designed sure-select beads and indexed for sequencing of 200 × 10<sup>6</sup> fragments per sample (100 bp paired-end) following the Agilent instructions. Capture Hi-C experiments were performed as biological duplicates.

Raw sequence reads of capture Hi-C (cHi-C) were mapped to the hg19 version of the human genome using BWA (v0.7.17) with parameters (mem -A 1 -B 4 -E 50 -L 0). Mapped reads were further processed by HiCExplorer (v3.5.1) to remove duplicated reads and reads from dangling ends, self-circle, self-ligation, and same fragments. The replicates of the same samples were compared, and confirmed to have high consistency (Pearson correlation coefficient: 0.83–0.99), then were merged to construct contact matrices of 2 kb resolution. Normalization was performed to ensure that all samples have the same number of total contacts, followed by KR normalization. The relative contact difference between two cHi-C maps was calculated by subtracting one from the other after scaling one sample to the other by using the total number of contacts in each sample.

**RNA-sequencing and data analysis.** Triplicates of either undifferentiated or differentiated ZIP13K2<sup>31</sup> cultures were treated with Accutase (Sigma-Aldrich, A6964) and differentiated cultures were further quenched with FACS buffer containing 5 mM EDTA (ThermoFischer Scientific, 15575020) 10% FBS (ThermoFischer Scientific, 26140079) in DPBS (Thermo Fisher Scientific, 14190250) to obtain single cells. In order to enrich for CXCR4<sup>-</sup> or CXCR4<sup>+</sup> cell fractions of differentiated cultures, cells were stained for anti-Human CXCR4 (CD184) PE (as described under 21. FACS) and compared to Isotype and unselected control sorted for either CXCR4<sup>-</sup> or CXCR4<sup>+</sup> subpopulations on the Aria II (Beckton Dickinson). RNA isolation including on-column DNase digest of enriched cell populations was performed using the RNeasy Mini Kit (Qiagen, 74104) according to the manufacturer's instructions. The KAPA Stranded mRNA-Seq Kit (Kapa Biosystems, #KK8401) was utilized for RNA library preparation, using 500 ng total RNA and performing poly-(A) enrichment followed by first-strand cDNA-synthesis (11 cycles). Subsequently, RNA-sequencing libraries were prepared by the use of dual index primers according to the manufacturer's instructions. Illumina adapter ligated sequencing libraries were sequenced for 50 million 75 bp long read pairs per sample on the HiSeq4000 (Illumina). RNA-seq data were preprocessed using cutadapt<sup>71</sup> to remove adapter sequences and trim low-quality bases. Reads were aligned against hg19 using STAR<sup>72</sup> (v 2.6.1d, parameter: -outSAMtype BAM SortedByCoordinate -outSAMattributes Standard -outSAMstrandField intronMotif -outSAMunmapped Within -quantMode GeneCounts). Subsequently, Stringtie<sup>73</sup> (v 1.3.5) was used for transcript assembly, e.g., calculation of strand-specific TPMs. Differential expression analysis was done independently per group comparison using the R package DESeq2<sup>74</sup> utilizing the raw expression counts from STAR's reads per gene output and filtered for an adjusted *p*-value < 0.05 and a log<sub>2</sub> fold-change > 1. The PCA was calculated on the log<sub>2</sub>+1 normalized TPMs of the 100 most variable genes using the R function pcomp (parameters "center = TRUE, scale = TRUE"). Box- and scatter plots show the unmodified TPMs. The heatmaps shows Z-score normalized TPMs to adjust for differences in absolute expression levels and was plotted using the R package pheatmap.

**qRT-PCR gene expression.** TaqMan-based qRT-PCR reactions were set up in triplicate using the 2 × TaqMan Fast Advanced Master Mix (Thermo, 4444557) according to manufacturer's instructions. Reactions were run on a StepOnePlus (Thermo) PCR machine with 40 cycles of 1 s at 95 °C and 20 s at 60 °C. following TaqMan probes (Thermo) were used: SOX17 Hs00751752\_s1; NANOG Hs02387400\_g1; T/BRACHYURY Hs00610080\_m1; FOXA2 Hs00232764\_m1; GATA4 Hs00171403\_m1; DKK1 Hs00183740\_m1; DKK2 Hs00205294\_m1; DKK4 Hs00205290\_m1; 18 s Hs03003631\_g1. NANOGNB Hs04225119\_g1; GDF3 Hs00220998\_m1; APOBEC1 Hs00242340\_m1, DPPA3 Hs01931905\_g1; CLEC4C Hs01092460\_g1; ATP6V1H Hs00977530\_m1; RGS20 Hs00991569\_m1; TCEA1 Hs04403253\_g1; LYPLA1 Hs00911024\_g1; MRPL15 Hs00204356\_m1; RP1 Hs00196698\_m1.

**ChIP qRT-PCR.** For CTCF-ChIP qRT-PCR, undifferentiated ZIP13K2<sup>31</sup> cells were grown to a final count of 10 million, treated with Accutase (Sigma–Aldrich, A6964), resuspended and washed in DPBS. Subsequently, cells were crosslinked in 1% formaldehyde solution for 5 min at room temperature. Following quenching with 0.125 M glycine final and two DPBS washes, we isolated nuclei using 1 ml cell lysis buffer (20 mM Tris-HCl pH8.0, 85 mM KCl, 0.5% NP40) for 10 min on ice. Then nuclei were spun down for 3 min at 2500 × g and supernatant was removed. The pellet was resuspended in 1 ml of nuclear lysis buffer (10 mM Tris-HCl, pH 7.5, 1% NP40, 0.5% sodiumdeoxycholate, 0.1% SDS) then incubated for 10 min on ice. Sonication was carried out on a Covaris E220 Evolution sonicator (PIP = 140.0, Duty Factor = 5.0, Cycles/Burst = 200, 10 min). After sonication, chromatin was spun down at 15,000 × g for 10 min to pellet insoluble material. Volume was increased to 1.5 mL with Chip Dilution Buffer (0.01% SDS, 1.1% Triton-X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.1, 167 mM NaCl), and 20 µl of CTCF antibody (CST, D31H2-XP) was added. Immunoprecipitation mixture was allowed to rotate overnight at 4 °C. The next day, 40 µl of Protein A Dynabeads (Thermo, 10001D) were added to the IP mixture and allowed to rotate for 4 h at 4 °C. This was followed by two washes of each: low salt wash buffer (0.1% SDS, 1% Triton-X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.1, 150 mM NaCl); high salt wash buffer (0.1% SDS, 1% Triton-X-100, 2 mM EDTA, 20 mM Tris, pH 8.1, 500 mM NaCl); LiCl wash buffer (0.25 M LiCl, 1% NP40, 1% deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.1); and TE buffer pH 8.0 (10 mM Tris-HCl, pH 8.0, 1 mM EDTA pH 8.0). DNA was eluted twice using 50 µl of elution buffer (0.5–1% SDS and 0.1 M NaHCO<sub>3</sub>) at 65 °C for 15 min. 16 µl of reverse crosslinking salt mixture (250 mM Tris-HCl, pH 6.5, 62.5 mM EDTA pH 8.0, 1.25 M NaCl, 5 mg/ml Proteinase K) was added and samples were allowed to incubate at 65 °C overnight. DNA was purified using AMPure XP beads (Beckman-Coulter) and treated with DNase-free RNase (Roche) for 30 min at 37 °C.

qRT-PCR reactions were set up in triplicate with the 2× PowerUp SYBR Green Master Mix (Thermo, A25742). Reactions were run on a StepOnePlus (Thermo) PCR machine with 40 cycles of 15 s at 95 °C and 60 s at 60 °C (Supplementary Table 6).

## Data availability

The data that support this study are available from the corresponding authors upon reasonable request. All Hi-C, RNA-seq, and capture Hi-C data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number [GSE127196](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127196). The Hi-C data used in this study are available in the GEO database under accession number [GSE52457](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52457) and [GSE63525](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525). Hi-C data of human embryos were obtained from the Genome Sequence Archive with the accession number [GSA0000852](https://www.genome-sa.org/GSA0000852). CTCF-ChIP-seq data used in this study are available at [Cistrome](https://www.cistrome.org/) (<http://www.cistrome.org>) and in the GEO database under accession number [GSM518375](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM518375), [GSM325899](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM325899), [GSM614637](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614637), [GSM614636](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614636), [GSM614631](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614631), [GSM614630](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614630), [GSM325899](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM325899), [GSM614615](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614615), [GSM614614](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM614614), [GSM651543](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM651543), [GSM651542](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM651542), [GSM651541](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM651541), [GSM586888](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM586888), [GSM586887](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM586887), [GSM534492](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM534492), [GSM534485](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM534485), [GSM534478](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM534478), [GSM534471](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM534471), [GSM325895](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM325895), [GSM489290](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489290), [GSM489291](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489291), [GSM489292](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489292), [GSM489293](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489293), [GSM489294](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489294), [GSM489295](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489295), [GSM489296](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489296), [GSM489297](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489297), [GSM489298](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489298), [GSM489299](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489299), [GSM489300](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489300), [GSM489301](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489301), [GSM489302](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM489302), [GSM782156](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM782156), [GSM782155](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM782155), [GSM631475](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631475), [GSM631476](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631476), [GSM631477](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631477), [GSM631478](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631478), [GSM631479](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631479), [GSM624077](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM624077), [GSM624078](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM624078), [GSM624079](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM624079), [GSM624080](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM624080), [GSM624081](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM624081), [GSM748538](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM748538), [GSM748539](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM748539), [GSM941710](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM941710), [GSM1056576](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1056576), [GSM1056577](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1056577), [GSM1070125](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1070125), [GSM646475](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646475), [GSM646474](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646474), [GSM646432](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646432), [GSM646433](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646433), [GSM1138985](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1138985), [GSM822276](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM822276), [GSM822277](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM822277), [GSM822278](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM822278), [GSM1007997](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1007997), [GSM1007998](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1007998), [GSM646373](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646373), [GSM646315](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646315), [GSM646334](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646334), [GSM646353](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646353), [GSM646372](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646372), [GSM646354](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646354), [GSM646314](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646314), [GSM646335](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646335), [GSM646392](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646392), [GSM646393](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM646393), [GSM808772](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808772), [GSM808773](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808773), [GSM808774](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808774), [GSM808775](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808775), [GSM808776](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808776), [GSM808777](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808777), [GSM808778](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808778), [GSM808779](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808779), [GSM808780](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808780), [GSM808781](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808781), [GSM808782](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808782), [GSM808783](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808783), [GSM808784](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808784), [GSM808785](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808785), [GSM808786](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808786), [GSM808787](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808787), [GSM808788](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808788), [GSM808789](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808789), [GSM808790](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808790), [GSM808791](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808791), [GSM808792](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808792), [GSM808793](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808793), [GSM808794](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808794), [GSM808795](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808795), [GSM808796](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808796), [GSM808797](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808797), [GSM808798](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808798), [GSM808799](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808799), [GSM808800](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM808800), [GSM1208603](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1208603), [GSM947527](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM947527), [GSM947528](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM947528), [GSM849300](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849300), [GSM849301](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849301), [GSM849302](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849302), [GSM849303](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849303), [GSM849304](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849304), [GSM849305](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM849305), [GSM970828](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM970828), [GSM1055825](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1055825), [GSM1224649](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224649), [GSM1224650](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224650), [GSM1224651](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224651), [GSM1224652](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224652), [GSM1224653](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224653), [GSM1224654](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224654), [GSM1224655](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224655), [GSM1224656](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224656), [GSM1224657](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224657), [GSM1224658](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224658), [GSM1224659](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224659), [GSM1224660](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1224660), [GSM1233869](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233869), [GSM1233870](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233870), [GSM1233871](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233871), [GSM1233872](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233872), [GSM1233873](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233873), [GSM1233874](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233874), [GSM1233875](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233875), [GSM1233876](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233876), [GSM1233877](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233877), [GSM1233878](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233878), [GSM1233879](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233879), [GSM1233880](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233880), [GSM1233914](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233914), [GSM1233915](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233915), [GSM1233916](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233916), [GSM1233917](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233917), [GSM1233918](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233918), [GSM1233919](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233919), [GSM1233920](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233920), [GSM1233921](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233921), [GSM1233922](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233922), [GSM1233923](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233923), [GSM1233924](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233924), [GSM1233925](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233925), [GSM1233926](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233926), [GSM1233927](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233927), [GSM1233928](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233928), [GSM1233929](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233929), [GSM1233930](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233930), [GSM1233931](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233931), [GSM1233932](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233932), [GSM1233933](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233933), [GSM1233934](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233934), [GSM1233935](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233935), [GSM1233936](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233936), [GSM1233937](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233937), [GSM1233938](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233938), [GSM1233939](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233939), [GSM1233940](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1233940), [GSM1234010](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234010), [GSM1234011](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234011), [GSM1234012](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234012), [GSM1234013](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234013), [GSM1234014](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234014), [GSM1234015](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234015), [GSM1234016](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234016), [GSM1234017](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234017), [GSM1234018](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234018), [GSM1234019](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234019), [GSM1234020](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234020), [GSM1234021](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234021), [GSM1234022](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234022), [GSM1234023](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234023), [GSM1234024](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234024), [GSM1234025](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234025), [GSM1234026](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234026), [GSM1234027](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234027), [GSM1234028](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234028), [GSM1234029](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234029), [GSM1234030](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234030), [GSM1234031](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234031), [GSM1234032](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234032), [GSM1234033](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234033), [GSM1234034](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234034), [GSM1234035](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234035), [GSM1234036](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234036), [GSM1234037](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234037), [GSM1234038](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234038), [GSM1234039](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234039), [GSM1234040](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234040), [GSM1234041](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234041), [GSM1234042](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234042), [GSM1234043](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234043), [GSM1234044](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234044), [GSM1234045](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234045), [GSM1234046](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234046), [GSM1234047](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234047), [GSM1234048](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234048), [GSM1234049](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234049), [GSM1234050](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234050), [GSM1234121](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234121), [GSM1234122](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234122), [GSM1234144](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234144), [GSM1234145](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234145), [GSM1234146](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234146), [GSM1234147](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234147), [GSM1234148](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234148), [GSM1234149](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234149), [GSM1234150](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234150), [GSM1234181](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234181), [GSM1234182](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234182), [GSM1234198](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234198), [GSM1234199](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234199), [GSM1234200](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234200), [GSM1234216](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234216), [GSM1234217](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234217), [GSM1234218](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234218), [GSM1234219](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1234219), [GSM1239390](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1239390), [GSM1239588](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1239588), [GSM1240813](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1240813), [GSM1240827](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1240827), [GSM1335528](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1335528), [GSM1003582](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1003582), [GSM1003581](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1003581), [GSM1003474](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1003474), [GSM1003464](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1003464), [GSM733752](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733752), [GSM733751](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733751), [GSM733750](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733750), [GSM733749](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733749), [GSM733748](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733748), [GSM733747](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733747), [GSM733746](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733746), [GSM733745](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733745), [GSM733744](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733744), [GSM733743](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733743), [GSM733742](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733742), [GSM733741](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733741), [GSM733740](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733740), [GSM733739](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733739), [GSM733738](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733738), [GSM733737](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733737), [GSM733736](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733736), [GSM733735](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733735), [GSM733734](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733734), [GSM733733](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733733), [GSM733732](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733732), [GSM733731](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733731), [GSM733730](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733730), [GSM733729](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733729), [GSM733728](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733728), [GSM733727](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733727), [GSM733726](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733726), [GSM733725](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733725), [GSM733724](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733724), [GSM733723](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733723), [GSM733722](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733722), [GSM733721](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733721), [GSM733720](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733720), [GSM733719](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733719), [GSM733718](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733718), [GSM733717](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733717), [GSM733716](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733716), [GSM733715](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733715), [GSM733714](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733714), [GSM733713](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733713), [GSM733712](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733712), [GSM733711](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM733711),

17. Kragestein, B. K. et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* **50**, 1463–1473 (2018).
18. Giorgio, E. et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum. Mol. Genet.* **24**, 3143–3154 (2015).
19. Redin, C. et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2017).
20. Ji, X. et al. 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).
21. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 e324 (2017).
22. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**, eaaw1668 (2019).
23. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
24. Rowley, M. J. et al. Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals. *Genome Res.* **30**, 447–458 (2020).
25. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
26. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
27. Chen, X. et al. Key role for CTCF in establishing chromatin structure in human embryos. *Nature* **576**, 306–310 (2019).
28. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014).
29. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.* **20**, 535–550 (2019).
30. D'Amour, K. A. et al. Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotechnol.* **23**, 1534–1541 (2005).
31. Tandon, R. et al. Generation of two human isogenic iPSC lines from fetal dermal fibroblasts. *Stem Cell Res.* **33**, 120–124 (2018).
32. Lu, L. et al. Robust Hi-C maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. *Mol. Cell* **79**, 521–534 e515 (2020).
33. Kanai-Azuma, M. et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. *Development* **129**, 2367–2379 (2002).
34. Tsankov, A. M. et al. Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–349 (2015).
35. Yang, J. et al. Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **21**, 341–352 (2020).
36. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
37. Pauklin, S. & Vallier, L. The cell-cycle state of stem cells determines cell fate propensity. *Cell* **155**, 135–147 (2013).
38. Mao, B. & Niehrs, C. Kremen2 modulates Dickkopf2 activity during Wnt/LRP6 signaling. *Gene* **302**, 179–183 (2003).
39. Engert, S. et al. Wnt/beta-catenin signalling regulates Sox17 expression and is essential for organizer and endoderm formation in the mouse. *Development* **140**, 3128–3138 (2013).
40. Sinner, D., Rankin, S., Lee, M. & Zorn, A. M. Sox17 and beta-catenin cooperate to regulate the transcription of endodermal genes. *Development* **131**, 3069–3080 (2004).
41. Mukherjee, S. et al. Sox17 and beta-catenin co-occupy Wnt-responsive enhancers to govern the endoderm gene regulatory network. *elife* <https://doi.org/10.7554/eLife.58029> (2020).
42. Banaszynski, L. A., Chen, L. C., Maynard-Smith, L. A., Ooi, A. G. & Wandless, T. J. A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules. *Cell* **126**, 995–1004 (2006).
43. Heidari, N. et al. Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).
44. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
45. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
46. de Wit, E. et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).
47. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
48. Mitsui, K. et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642 (2003).
49. Despang, A. et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
50. Fei, T. et al. Deciphering essential cisomes using genome-wide CRISPR screens. *Proc. Natl Acad. Sci. USA* **116**, 25186–25195 (2019).
51. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.* **20**, 2349–2354 (2006).
52. Brown, J. M. et al. A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nat. Commun.* **9**, 3849 (2018).
53. Deng, W. et al. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849–860 (2014).
54. Kim, J. H. et al. LADL: light-activated dynamic looping for endogenous gene expression control. *Nat. Methods* **16**, 633–639 (2019).
55. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
56. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
59. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
60. Liu, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
61. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
62. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
63. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
64. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
65. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
66. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
67. Zhou, X. et al. Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods* **10**, 375–376 (2013).
68. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
69. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
70. Yin, Y. et al. Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. *Cell Stem Cell* **16**, 504–516 (2015).
71. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

### Acknowledgements

We would like to thank the Michor and Meissner labs for helpful discussions as well as Bernd Timmermann and the MPIMG sequencing core facility for the ongoing support. We gratefully acknowledge support of the Dana-Farber Cancer Institute Physical Sciences-Oncology Center, NIH U54CA193461 (to F.M.), NIH 1P50HG006193, P01GM099117, 1DP3K111898 (to A.M.), and the Max Planck Society (to A.M.). We acknowledge support of the High-performance Computing Platform of Peking University.

### Author contributions

H.J.W. conceived, planned, designed, and performed the data analyses. A.L. conceived, planned, designed, and performed the experiments. E.S. and A.B. helped in performing experiments. H.K. performed the bioinformatics analysis of RNA seq. A.M. and F.M. conceptualized, designed, and supervised the study. H.J.W., A.L., A.M., and F.M. wrote the manuscript. All authors contributed to the final paper.

### Competing interests

A.M. and F.M. are co-founders of an oncology company. The remaining authors declare no competing interests.