



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

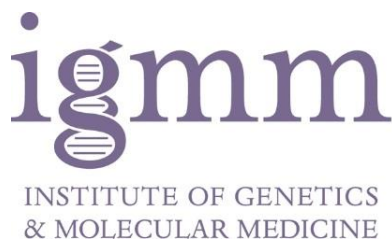
The relationship between chromatin structure and distal regulation

Lauren Kane

Thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2021



Declaration

This thesis has been composed entirely by me. Apart from where stated otherwise, the experiments were carried out solely by me. The work has not been submitted for any other degree or professional qualification.

Lauren Kane

March 2021

Acknowledgements

Firstly, to my PhD supervisors Bob Hill and Wendy Bickmore, thank you both for providing me with the opportunity to carry out my research across both labs. I am very grateful for all the encouragement, advice, critique and support I have received from you both for the duration of my PhD. I am so glad I was able to be part of both lab groups and even managed to pretty much keep a 50/50 split!

To Laura and Iain – let's be honest I may not have survived without you both! Thank you for all the experimental training, moral support, advice and most importantly, lots of laughs. I am very grateful you allowed me to join the CTCF family! Although I am pretty sure you both said here is something to get you started whilst we work out your 'real project'. I will miss working with you lots!

A huge thank you to the current and previous members of both the Hill and Bickmore lab groups. I think almost every single person has helped me out with some of experimental or life dilemma at some point. A special thank you to Shelagh Boyle for lots of technical expertise, life advice and ensuring there was always a steady flow of baked goods at coffee breaks.

I am very grateful to everyone in the HGU, especially the Genome Regulation and Disease Mechanisms sections – it was great to be able to do my PhD in such a friendly and social environment. I have made so many friends across the institute and I am very sad to say goodbye. Special thanks to all the technical services here at the IGMM. I spent a LOT of my PhD down in microscopy trying to make sense of spots so a big thank you to the AIR team especially Matt who put up with many stupid questions (a potentially broken microscope) and Laura who was always around for a natter as well as support with any analysis. Also, thank you to Lizzie and the rest of the FACS team – it was always a pleasure to come down and visit you.

To the 2016 cohort; Kathryn, Lloyd, Emma, Juliet, Becka, Liusaidh, Carolin, Nikki, Christina and Imke, I could not have asked for a better group of friends to go through my PhD journey with and I am thankful for all your support, fun times and crossword

lunchtimes. To Zoe, Abby, Gabi and Luke – I am still pretty annoyed you all left me but I am grateful that I had you for my office, lab and going out buddies! To my fellow legacy players Kathryn, Matthew and Ilya – thank you for many, many nights of games, food, wine and lots of laughs. Ilya hopefully soon we can continue our end of the night scientific discussions (I think this was when I had all my best ideas??). To Beth, my oldest and closest friend, I am so thankful for all your support and understanding – I cannot wait until we next get away and celebrate all the new beginnings. Many thanks to my other ‘from home’ friends – Paul, Heather, Amy and Charlie.

I would like to thank my siblings, Claire, Nicole, Alfie and Dexter and my niece Millie – you never fail to put a smile on my face. To the rest of my family, I am so grateful for you all and the belief you have in me.

Finally I would like to thank my parents, Jacqui and Paul. You always go above and beyond to provide me with so much support. I was fortunate enough to always be encouraged by you both to just ‘do what you enjoy and find interesting’. Hopefully I will keep this up for the remainder of my career.

Abstract

Mammalian genomes are organised into topologically associated domains (TADs) and chromatin loops through a loop extrusion process. Structuring of chromatin within these highly conserved TADs is believed to both facilitate gene regulation via distal enhancers whilst insulating their regulatory activity at TAD boundaries. *Shh* and its regulatory elements are all contained within a highly conserved, well-characterised 1 Mb TAD that has previously been proposed to form an invariant chromatin loop structure across various tissue types. The overall aim of my thesis is to further understand how distal enhancers operate to regulate their target gene during mammalian development. To do this, I used the *Shh* as a model locus to study the spatiotemporal dynamics of distal enhancers and the mechanisms of long-range gene regulation in relation to three-dimensional chromatin structure.

To explore the spatiotemporal regulation of *Shh* across brain development I first performed ATAC-Seq on dissected, FAC-sorted embryonic brain tissue. I show that the *Shh* brain enhancers are accessible across embryonic brain development and that the *Shh* forebrain enhancer, SBE2, shows increased spatial proximity with the *Shh* gene specifically in SBE2-regulated cells. Next, to investigate the importance of TAD integrity on developmental gene regulation, we have manipulated the *Shh* TAD by deleting CTCF sites at the TAD boundaries. Both RNA and DNA fluorescence in situ hybridisation assays were used to investigate changes in *Shh* expression and chromatin conformation that result from these manipulations in single cells of developing embryonic tissue. I found that *Shh* loses its co-localisation with the tissue-specific enhancers in the CTCF mutant embryos but with no consequence on *Shh* expression patterns. I then used degron-tagged CTCF and cohesin embryonic stem cell lines to investigate if genome-wide depletion of these TAD architectural proteins affected *Shh* distal regulation using synthetic transcription factors. I found that cohesin is required for enhancer-driven gene activation, whereas CTCF was dispensable. Finally, I used patient-derived cell lines to study a human-specific *SHH* TAD boundary deletion found in acheiropodia patients. I show using DNA fluorescence in situ hybridisation

that this deletion causes ectopic interactions between the *SHH* TAD boundary and a novel CTCF peak detected in patient cells, therefore, potentially preventing ZRS from regulating the promoter.

Overall, my data shows that the removal of single CTCF sites in mice alters TAD structure but has no readily detectable effect on *Shh* expression patterns during development and results in no evident phenotypes. This suggests that the developmental regulation of *Shh* expression is remarkably robust to TAD perturbations. I then go on to show that distal regulation is dependent on cohesin but not CTCF. I propose a model where cohesin extrusion brings together distal elements to induce expression and CTCF functions to stabilise and insulate these interactions. This supports the view that CTCF provides robustness to developmental gene regulation. Ectopic CTCF binding in acheiropodia patients would block extrusion of cohesin and prevent *SHH*/ZRS interaction during embryonic development.

Lay summary

Birth defects occur in approximately 3% of births and are defined as structural changes identified at birth that can affect different parts of the body. For many of these defects the underlying cause is still unknown. Mutations in genes that underly these birth defects have been identified but for some patients mutations in DNA sequences that do not contain genes have been found. One example of these other DNA sequences is called enhancers which function to switch on genes. The aim of my thesis is to understand how genes are switched on by enhancers at critical points of development. Often these enhancers can be found at a great distance from the gene that they control, and it is not well understood how they exert their function over long-range. One feature of DNA that has been linked to long-range enhancer function is the non-random folding of DNA inside the nucleus. DNA folding into specific structures is thought to bring these long-range enhancers closer to their target gene and prevent them from accessing off-target genes. Some of the key proteins implicated in DNA folding are cohesin and CTCF. Cohesin is a ring-shaped protein that can trap DNA and move along it forming loops whereas CTCF is a protein that binds DNA CTCF can block cohesin movement along DNA and prevents enhancers from accessing non-target genes. As a model system, I used the *Shh* gene to study how these long-range enhancers work. *Shh* is a key gene involved in the development of many different tissues including limb, central nervous system, craniofacial elements, gut and lungs and has been implicated in several birth defects. *Shh* is switched on by different tissue-specific long-range enhancers during embryonic development.

To investigate the link between DNA folding and long-range enhancer function we removed single CTCF binding sites from the DNA surrounding the *Shh* gene and its enhancers in mice. I show using imaging of single-cells in embryonic tissue that CTCF site deletions disrupts the DNA structure between the *Shh* gene and its tissue-specific long-range enhancers, but these enhancers are still able to switch *Shh* on. I next used cell lines that were made from early embryos to test whether *Shh* could be switched on from its long-range enhancers using artificial activator proteins in the absence of either

all CTCF or cohesin protein in the cells. I found that cohesin was required for long-range enhancer-driven gene activation whereas CTCF was dispensable. Finally, I used patient-derived cell lines to study how removal of a human-specific CTCF binding site cluster is related to the genetic condition acheiropodia. I show by imaging of these cell lines that loss of this CTCF cluster causes a new CTCF cluster to change the DNA folding and consequently blocks one of the long-range enhancers from switched on *Shh*.

Overall, my data shows that removal of single CTCF sites can change DNA structure but has no effect on the function of long-range enhancers. This suggests that other compensatory mechanisms may be in place to ensure *Shh* is correctly switched on at critical developmental windows. I then go on to show that long-range enhancer function is dependent of cohesin but not CTCF. Taking all this together, I propose that cohesin is required for long-range enhancers to switch on their target gene by bringing them together in the nucleus whereas CTCF is important for stabilising but not initiating these interactions.

Table of contents

Declaration	iii
Acknowledgments	v
Abstract	vii
Lay summary	ix
Table of contents	xi
List of figures	xvii
List of tables	xix
Abbreviations.....	xx

Chapter 1. Introduction..... 1-1

1.1 Mammalian genome organisation	1-3
1.1.1 Chromatin is organised in the nucleus across various scales	1-3
1.1.1.1 Chromosome territories	1-5
1.1.1.2 A/B Compartments	1-6
1.1.1.3 Topologically associated domains and chromatin loops.....	1-8
1.1.2 Loop extrusion produces chromatin loops and TADs	1-11
1.1.3 TADs as a functional unit of gene regulation.....	1-18
1.1.3.1 Correlation between TADs and regulatory landscapes.....	1-18
1.1.3.2 TAD disruption in human disease	1-20
1.2 Gene regulation during development.....	1-23
1.2.1 An overview of developmental enhancers	1-23
1.2.1.1 Enhancer have distinct chromatin features	1-23
1.2.1.2 Enhancers can behave additively or redundantly.....	1-24
1.2.2 Mechanisms of long-range enhancer regulation.....	1-25
1.2.2.1 Tracking and linking models	1-26
1.2.2.2 Looping model.....	1-26
1.2.2.3 Phase condensate model	1-28
1.2.2.4 Comparison of models.....	1-30
1.2.3 Enhancer dysregulation in human disease.....	1-32

1.3	<i>Shh</i>: A model locus of gene regulation.....	1-34
1.3.1	<i>Shh</i> in embryonic development	1-34
1.3.1.1	Identification of the <i>hedgehog</i> genes.....	1-34
1.3.1.2	<i>Shh</i> is a vertebrate morphogen.....	1-35
1.3.2	SHH signalling pathway during development	1-36
1.3.3	<i>Shh</i> has a complex regulatory landscape.....	1-39
1.3.3.1	Limb bud expression	1-39
1.3.3.2	Central nervous system expression.....	1-41
1.3.3.3	Epithelial tissue expression	1-43
1.3.4	Do the <i>Shh</i> enhancers have an enhancer element signature?	1-44
1.3.5	<i>Shh</i> functions in a three-dimensional genome.....	1-45
1.3.6	SHH misregulation can cause human developmental disorders.....	1-47
1.4	Aims of PhD thesis.....	1-50
 Chapter 2. Materials & Methods		2-51
2.1	General molecular biology reagents.....	2-53
2.2	Cell culture	2-54
2.2.1	Mouse embryonic stem cells.....	2-54
2.2.2	Human lymphoblastoid and fibroblast cell lines	2-54
2.3	Mouse tissue	2-55
2.3.1	Embryonic tissue dissection.....	2-55
2.3.2	Genotyping.....	2-55
2.3.3	Tissue embedding and sectioning	2-56
2.4	Assay for transposase-accessible chromatin.....	2-56
2.4.1	Mouse model information.....	2-56
2.4.2	Preparation of single-cell suspensions	2-57
2.4.3	Fluorescent-activated cell sorting	2-57
2.4.4	Transposition reaction.....	2-57
2.4.5	ATAC library amplification.....	2-58
2.4.6	ATAC library purification	2-60
2.4.7	ATAC library quality control and sequencing	2-60
2.5	ATAC-seq computational analysis.....	2-61
2.5.1	Read mapping	2-61

2.5.2	Quality control	2-61
2.5.3	Peak calling and analysis	2-61
2.6	Immunofluorescence.....	2-62
2.6.1	Immunofluorescence solutions.....	2-62
2.6.2	Immunofluorescence of mouse embryonic tissue	2-62
2.7	DNA fluorescent in situ hybridisation	2-63
2.7.1	Fosmid probe preparation	2-63
2.7.2	Fosmid probe labelling.....	2-64
2.7.3	3D FISH of tissue sections	2-65
2.7.4	3D FISH of cell lines.....	2-66
2.8	RNA fluorescent in situ hybridisation	2-67
2.8.1	RNA FISH of tissue sections	2-67
2.8.2	RNA FISH of cell lines	2-68
2.9	Imaging.....	2-68
2.9.1	Image acquisition and deconvolution.....	2-68
2.9.2	DNA FISH spot measurements	2-68
2.9.3	RNA FISH signal quantification	2-69
2.10	Synthetic biology tools	2-69
2.10.1	TALE design and assembly.....	2-69
2.10.2	mESC transfections	2-73
2.10.3	Auxin-inducible degron induction.....	2-73
2.10.4	Expression analyses.....	2-73
2.11	Statistical tests	2-74
 Chapter 3. <i>Shh</i> regulation during neural development.....		3-75
3.1	Introduction	3-77
3.2	ATAC-seq of <i>Shh</i>-expressing neural cells across development	3-78
3.2.1	Generation of ATAC-seq libraries from genetic mouse models	3-78
3.2.2	QC Analysis of ATAC-seq libraries	3-81
3.2.3	Chromatin accessibility in neural tissue is unaffected by <i>Shh</i> of developmental stage..	3-87
3.2.3.1	ATAC-seq profile of the <i>Shh</i> regulatory landscape	3-87
3.2.3.2	ATAC-seq profiles of other regulatory landscapes	3-91

3.2.4	Analysis of peaks unique to <i>Shh</i> -expressing cell populations.....	3-93
3.2.5	Motif analysis across developmental stages.....	3-96
3.2.5.1	Commonly occurring motifs.....	3-96
3.2.5.2	Motifs exclusive to cell type or developmental stage	3-97
3.3	<i>Shh</i> forebrain expression is distally regulated	3-99
3.3.1	Chromatin conformation of SBE2-regulated forebrain tissue.....	3-100
3.3.2	Development of an RNA FISH assay to measure the proportion of expressing alleles in embryonic tissue.....	3-107
3.3.3	Expression analysis of SBE2 regulated forebrain tissue	3-109
3.4	Discussion & conclusions	3-111
3.4.1	Use of a <i>Shh</i> reporter mouse model to study enhancer regulation	3-111
3.4.2	Chromatin accessibility marks enhancers but not their activity	3-112
3.4.3	Identification of temporally-regulated transcriptional networks in <i>Shh</i> -expressing cells from neural tissue.....	3-113
3.4.4	Assessment of the robustness of ATAC-seq datasets	3-114
3.4.5	Distal forebrain enhancer of <i>Shh</i> is found in close proximity	3-115
3.4.6	SBE7-mediated gene regulation in the developing forebrain.....	3-116
Chapter 4.	The role of CTCF in <i>Shh</i> distal regulation	4-117
4.1	Introduction	4-119
4.1.1	CTCF binding sites deletion alters <i>Shh</i> TAD structure in mESCs	4-120
4.1.1.1	<i>Shh</i> TAD boundary.....	4-120
4.1.1.2	<i>Lmbr1</i> TAD boundary	4-121
4.2	Tissue-specific enhancer-promoter proximity is lost in ΔCTCF mutants	4-124
4.2.1	<i>Shh</i> -SBE2 in forebrain tissue	4-124
4.2.2	<i>Shh</i> -ZRS in limb tissue	4-125
4.3	<i>Shh</i> developmental expression patterns are unaffected in ΔCTCF mutants	4-129
4.3.1	Δ CTCF mutants display normal <i>Shh</i> expression patterns	4-129
4.3.2	<i>Shh</i> transcriptional levels are within the normal range	4-130
4.4	Discussion & conclusions	4-132
4.4.1	Developmental gene regulation of <i>Shh</i> is not solely determined by CTCF-mediated interactions	4-132
4.4.2	Approaches to study chromatin organisation	4-133

4.4.3	CTCF binding sites provide transcriptional robustness.....	4-134
4.4.4	Is direct interaction between enhancers and promoters required?	4-135

Chapter 5. Synthetic approach to study *Shh* regulation 5-137

5.1	Introduction	5-139
5.2	<i>Shh</i> expression is induced by synthetic activation of distal enhancers.....	5-140
5.2.1	Construction of tZRS-VP64.....	5-140
5.2.2	Validation of synthetic transcription factors	5-143
5.3	Acute degradation of architectural proteins disrupts <i>Shh</i> TAD structure.....	5-148
5.3.1	Degron-tagged CTCF and SCC1 are depleted upon auxin treatment.....	5-148
5.3.2	CTCF or SCC1 depletion disrupts the <i>Shh</i> TAD	5-151
5.3.3	Auxin treatment does not affect TALE activity in mESCs	5-153
5.4	Distal activation of <i>Shh</i> does not depend on CTCF	5-155
5.5	Cohesin is necessary for distal activation of <i>Shh</i>	5-157
5.6	Discussion & conclusions	5-159
5.6.1	Use of synthetic transcription factors to study gene regulation.....	5-159
5.6.2	Cohesin is required for enhancer-driven gene activation	5-161
5.6.3	CTCF is indirectly involved in gene regulation	5-163

Chapter 6. *SHH* TAD structure in humans.....6-165

6.1	Introduction	6-167
6.1.1	Acheiropodia causes severe limb defects.....	6-167
6.1.2	Acheiropodia patients have a Δ 12 kb deletion that encompasses CTCF binding sites	6-167
6.1.3	Comparison of the human <i>SHH</i> TAD to mouse.....	6-168
6.2	Analysis of <i>SHH</i> TAD structure in human cell lines	6-173
6.2.1	DNA FISH of acheiropodia patient-derived cell lines	6-173
6.2.2	Allele-specific DNA FISH of patient and parental cell lines	6-175
6.2.3	<i>SHH</i> TAD structure in primary human fibroblasts.....	6-177
6.3	Discussion & conclusions	6-181
6.3.1	Removal of CTCF sites causes acheiropodia in humans.....	6-181
6.3.2	Δ 12 kb mutation specifically affects ZRS long-range regulation.....	6-182

6.3.3	Using differentiated cell lines to model human disease	6-183
6.3.4	Comparison of the mouse and human <i>SHH</i> TAD	6-184
Chapter 7.	Concluding remarks.....	7-185
Chapter 8.	Appendices	8-191
8.1	Additional Figures	8-193
8.2	Publication	8-199
References		8-201

List of Figures

Figure 1.1. Chromatin is organised in the nucleus across various scales.	1-11
Figure 1.2. Loop extrusion produces topologically associated domains and chromatin loops.	1-14
Figure 1.3 Depletion of CTCF or cohesin as a means to study TAD functions.	1-18
Figure 1.4. Human structural variants disrupt TADs and can have pathological consequences due to enhancer adoption.	1-22
Figure 1.5. There are a number of models of long-range enhancer regulation.	1-30
Figure 1.6. <i>Shh</i> is broadly expressed during embryonic development.	1-36
Figure 1.7. The SHH signalling pathway.	1-38
Figure 1.8 The <i>Shh</i> regulatory landscape.	1-40
Figure 1.9. <i>Shh</i> regulation in neural tissue.	1-42
Figure 2.1. TALE construct maps.	2-72
Figure 3.1 Isolation of <i>Shh</i> -expressing neural cells for ATAC-seq.	3-79
Figure 3.2 Library preparation of transposed samples for sequencing.	3-80
Figure 3.3 QC analysis of E10.5 ATAC samples following sequencing.	3-82
Figure 3.4 PCA analysis of ATAC samples.	3-84
Figure 3.5 IDR analysis of ATAC samples.	3-86
Figure 3.6 Chromatin accessibility across the <i>Shh</i> regulatory landscape.	3-89
Figure 3.7 <i>Shh</i> neural enhancers and CTCF sites can be detected by ATAC-seq.	3-90
Figure 3.8 Chromatin accessibility profiles of other gene regulatory landscapes.	3-92
Figure 3.9 Identification of unique accessible regions in <i>Shh</i> -expressing cells.	3-94
Figure 3.10 Genomic distribution and gene ontology analysis of unique peaks from <i>Shh</i> -expressing cells.	3-95
Figure 3.11 Motif analysis of across neural development.	3-98
Figure 3.12 Identification of the SBE2-regulated region of the forebrain.	3-102
Figure 3.13 <i>Shh</i> and its forebrain enhancer show increased proximity in SBE2 active regions of the forebrain.	3-104
Figure 3.14 The <i>Shh</i> TAD is more compact in <i>Shh</i> -expressing tissue.	3-106
Figure 3.15 Visualising <i>Shh</i> nascent expression using oligo probes.	3-108

Figure 3.16 <i>Shh</i> is expressed at similar levels in the forebrain across developmental stages.	3-110
Figure 4.1 Removal of CTCF binding sites at the <i>Shh</i> TAD boundaries.	4-123
Figure 4.2 SBE2 shows increased distance from the <i>Shh</i> gene in the forebrain of Δ CTCF mutant embryos.	4-126
Figure 4.3 Decreased ZRS co-localisation with the <i>Shh</i> gene in the posterior limb bud of Δ CTCF mutant embryos.	4-127
Figure 4.4 <i>Shh</i> nascent transcription is largely unperturbed following removal of CTCF sites.	4-131
Figure 5.1 Design and assembly of a TALE-VP64 targeting the distal limb enhancer ZRS.	5-143
Figure 5.2 Nascent RNA FISH of <i>Shh</i> and <i>Lmbr1</i>	5-144
Figure 5.3 <i>Shh</i> expression can be induced using TALE-VP64 proteins targeting the <i>Shh</i> promoter or its distal enhancers.	5-146
Figure 5.4 Auxin treatment leads to targeted degradation of AID-tagged proteins.	5-150
Figure 5.5 Depletion of CTCF or SCC1 disrupts the <i>Shh</i> TAD structure.	5-152
Figure 5.6 Auxin treatment does not affect TALE-mediated <i>Shh</i> induction.	5-154
Figure 5.7 TALE-mediated <i>Shh</i> activation is independent of CTCF binding.	5-156
Figure 5.8 <i>Shh</i> activation using TALE-VP64 proteins targeting distal enhancers is dependent on cohesin.	5-158
Figure 6.1 Chromatin organisation of the human <i>SHH</i> TAD.	6-170
Figure 6.2 Comparison of the <i>Shh</i> TAD in mouse and humans.	6-171
Figure 6.3 Summary of mouse and human CTCF binding sites at the <i>Shh</i> locus.	6-172
Figure 6.4 Removal of 12 kb region of the <i>SHH</i> TAD alters CTCF looping interactions in patient-derived lymphoblastoid cell lines.	6-174
Figure 6.5 Altered CTCF looping interactions are specific to the chromosome carrying the 12 kb deletion.	6-177
Figure 6.6 Primary human fibroblasts differ from lymphoblastoid cell lines in their interactions with the <i>LMBR1</i> end of the <i>SHH</i> TAD.	6-179
Figure 7.1 The relationship between chromatin structure and distal regulation. ..	7-189

List of Tables

Table 2.1 General molecular biology reagents used.	2-53
Table 2.2 Genotyping thermal cycler programme.	2-55
Table 2.3. Nextera PCR primers used for library amplification.	2-59
Table 2.4. Library amplification thermal cycler programme.	2-59
Table 2.5. qPCR of amplified libraries thermal cycler programme.	2-60
Table 2.6. Immunofluorescence solutions.	2-62
Table 2.7. Details of fosmid probes.	2-64
Table 2.8. dUTPs used to label fosmid probes.	2-65
Table 2.9. Antibodies used to detect indirectly labelled fosmid probes.	2-66
Table 2.10. Golden-gate assembly of TALEs thermal cycler programme.	2-71
Table 2.11. Colony PCR of TALEs thermal cycler programme.	2-71
Table 2.12 TALE target DNA sequence.	2-71
Table 2.13. Gene expression analysis thermal cycler programme.	2-74

Abbreviations

3D	Three-dimensional
AID	Auxin-inducible degron
ATAC	Assay for transposase-accessible chromatin
Bp	Base pairs
BSA	Bovine serum albumin
CES	conserved essential surface
ChIP	Chromatin immunoprecipitation
CO ₂	Carbon dioxide
CTCF	CCCTC-binding factor
ddH ₂ O	Double distilled H ₂ O
DAPI	4',6-diamidino-2-phenylindole
DEPC	Diethyl pyrocarbonate
DISP1	Dispatched 1
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
ENCODE	Encyclopedia of DNA elements
EtOH	Ethanol
FACS	Fluorescence activated cell sorting
FCS	Foetal calf serum
FISH	Fluorescent in situ hybridisation
GC	Guanine cytosine
GFP	Green fluorescent protein
GLI	Glioma-associated
GO	Gene ontology

GWAS	Genome-wide association study
HCl	Hydrogen chloride
HPE	Holoprosencephaly
IDR	intrinsically disordered region
LADs	Lamina associated domains
LCR	Locus control region
LIF	Leukaemia inhibitory factor
LLPS	Liquid-liquid phase separation
Lmbr1	Limb region 1
LSC	LMBR1 – SHH CTCF
MACS	Mammal amphibian conserved sequence
mESCs	Mouse embryonic stem cells
MFCS	Mammal fish conserved sequence
MGE	Medial ganglionic eminence
MgCl ₂	Magnesium chloride
MgSO ₄	Magnesium sulphate
MRCS	Mammal reptile conserved sequence
mTmG	Membrane-targeted tdTomato/ membrane-targeted EGFP
NaCl	Sodium chloride
NaOH	Sodium Hydroxide
NH ₄ Cl	Ammonium chloride
O ₂	Oxygen
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PCS	Placental conserved sequence
PFA	Paraformaldehyde
PrCP	Prechordal plate
P/S	Penicillin/streptomycin
PTCH1	Patched 1
qPCR	Quantitative PCR

RNA	Ribonucleic acid
SBE	Shh brain enhancer
SDS	Sodium dodecyl sulphate
SFPE	Shh floorplate enhancer
Shh	Sonic hedgehog
SLGE	Shh lung and gut enhancer
SMO	Smoothened
SNP	Single nucleotide polymorphism
SSC	Saline sodium citrate
TAD	Topologically associated domain
TAE	Tris acetate EDTA
TALE	Transcription activator-like effector
TE	Tris EDTA
WISH	Wholemound in situ hybridisation
ZLI	Zona limitans intrathalamica
ZPA	Zone of polarising activity
ZRS	ZPA regulatory sequence

Chapter 1. Introduction

1.1 Mammalian genome organisation

Mammalian spatial genome organisation arises from the three-dimensional (3D) folding of chromatin within the nucleus. The specific arrangement of nuclear architecture is believed to be fundamental for genome function. The nucleus is not a fixed structure, rather its contents are dynamic with constant changes to chromatin structures that are necessary for essential cellular processes such as gene expression.

Below, I will discuss the different levels of chromatin organisation that have been observed inside the nucleus with a focus on chromatin loops, topologically associated domains (TADs), the mechanism that produces these structures and their function in the context of gene regulation.

1.1.1 Chromatin is organised in the nucleus across various scales

Mammals contain approximately 6 billion base pairs (bp) of DNA inside of each cell. This equates to around 2 metres of linear DNA that must be packed inside a nucleus that is typically 6 μm in diameter. This means our DNA must be highly compacted whilst remaining accessible in order for our genome to carry out its important functions such as DNA replication, recombination, repair and transcription.

At the smallest scale, the most basic unit of genome organisation is the nucleosome. Here, a canonical histone octamer that consists of two copies of the four core histone proteins (H2A, H2B, H3 and H4) is wrapped with 147 bp (1.7 turns) of DNA forming a nucleoprotein structure termed chromatin (Figure 1.1A) (Luger et al., 1997). Chromatin helps to compact DNA inside the nucleus whilst protecting it from physical or chemical damage and also facilitates DNA with carrying out its fundamental functions. Between nucleosomes is 20-40 bp of spacer DNA that is bound by linker histone H1 which functions to stabilise higher order structures (Woodcock et al., 2006). Nucleosomes are arranged into a 'beads-on-a-string' or 10 nm fibre structure (Finch and Klug, 1976). Depending on the function of a particular region of DNA, the histone proteins can be post-translationally modified at their core or at their tails (discussed further in 1.2.1) and can also be exchanged for different histone variants (Henikoff and Smith, 2015; Kouzarides, 2007). For example, regulatory regions typically contain the histone variants H2.AZ and H3.3. Modification to histone tails also function in the

compaction or decompaction of nucleosomes e.g. acetylation of particular lysines on the histone tail can decompact nucleosomes (Kouzarides, 2007). Nucleosome positioning along DNA is highly dynamic and ATP-dependent chromatin remodellers within the cell function to slide, assemble and disassemble nucleosome structures (Narlikar et al., 2013).

How the 10 nm fibre of chromatin is subsequently organised is somewhat controversial and is still under much debate. Evidence from *in vitro* studies of reconstituted purified nucleosomes using X-ray and Electron Microscopy (EM) suggesting that chromatin is tightly packed into 30 nm fibres that results from the supercoiling of nucleosomes and stabilised by linker histones and interactions between histone tails (Dorigo et al., 2004; Finch and Klug, 1976; Kruithof et al., 2009; Leth Bak et al., 1977; Robinson et al., 2006; Song et al., 2014; Widom and Klug, 1985). Using super-resolution microscopy to visualise nucleosomes *in vivo* in both fixed and live cells, it was shown more recently that rather than a regular fibre structure, nucleosomes were found in domains of varying size (Nozaki et al., 2017; Ricci et al., 2015). Interestingly, larger domains corresponded with heterochromatic regions whereas smaller domains correlated to regions of active transcription, indicative of a relationship between nucleosome organisation and function (discussed further in 1.1.1.2) (Ricci et al., 2015). More recent work which involved a modified type of EM to visualise chromatin called chromEMT found no evidence of a 30 nm fibre but instead saw a variety of polymorphic structures ranging from 5-24 nm inside mammalian cells (Figure 1.1B) (Ou et al., 2017). Altogether, this work argues against the textbook model of nucleosomes folding into a 30 nm fibre and then subsequent larger fibre structures and instead suggests that chromatin adopts a more flexible and disordered structure at this intermediate level of folding that can vary depending on the activity of that genomic region.

At larger scales (multi-kilobase to megabase pairs), chromatin is known to be organised into chromatin loops, TADs, compartments and chromosome territories. The main methods used to study these structures include microscopy-based techniques such as fluorescence *in situ* hybridisation (FISH), Oligopaint (Beliveau et al., 2012)

and EM, genomics-based approaches including Hi-C (Lieberman-Aiden et al., 2009), GAM (Beagrie et al., 2017), SPRITE (Quinodoz et al., 2018) and DamID (Van Steensel and Henikoff, 2000) and polymer-based modelling. Here, I will discuss each of these levels of chromatin organisation, the mechanisms driving their formation and how genome folding is intertwined with genomic functions.

1.1.1.1 Chromosome territories

In interphase nuclei, chromosomes are found arranged in a non-random manner occupying a defined nuclear volume termed a chromosome territory (reviewed by e.g. Cremer and Cremer, 2010). This organisation of chromosomes into distinct territories has been observed by DNA FISH using chromosome paints and later by Hi-C methods (Figure 1.1C) (Boyle et al., 2001; Lieberman-Aiden et al., 2009). Chromosomes also display a characteristic radial positioning within the nucleus with more gene-rich chromosomes found clustered together in the centre of the nucleus and gene-poor chromosomes found closer to the nuclear periphery (Figure 1.1C) (Bolzer et al., 2005; Boyle et al., 2001; Croft et al., 1999; Lieberman-Aiden et al., 2009). Genes have the ability to move out of their chromosome territory. For example, upon differentiation the *Hoxb* locus was shown to become decondensed and extrude out of its territory concurrently with the induction of gene expression (Chambeyron and Bickmore, 2004). In contrast, another Hox gene, *Hoxd*, does not need to loop out of its territory in all cell types during embryonic development for gene activation (Morey et al., 2007). Other work using the β -globin locus found that this region is located at nuclear periphery within its chromosome territory when the gene is repressed (Ragoczy et al., 2006). Upon erythroid differentiation, the β -globin moves out of its territory, into the interior of the nucleus as the gene becomes activated and this relocation is dependent on the Locus Control Region (LCR), a *cis*-acting regulatory sequence (Ragoczy et al., 2006). These studies suggested that, for at least some genes, the re-organisation of chromatin structure within a chromosome territory to re-position genes into a more transcriptionally active region of the nucleus may be important way in which to regulate gene expression. Another study used RNA FISH to track to position of expressing alleles and found that alleles outside their territory were no more likely to be transcriptionally active than alleles inside the chromosome territory (Morey et al.,

2009). The authors propose instead that increased chromatin mobility of genes is important for transcriptional activation as it increases the likelihood of interaction with transcriptional factories rather than the position of the gene within the chromatin territory.

1.1.1.2 A/B Compartments

Using DNA FISH, it was initially observed that within the same chromosome, gene clusters spatially segregate from gene deserts inside the nucleus (Shopland et al., 2006). There has been other studies demonstrating that active euchromatic domains of chromatin are found segregated from inactive heterochromatin (reviewed by e.g. Bickmore and Van Steensel, 2013). At the sub-chromosomal level, chromatin is partitioned into regions that preferentially interact with one another, termed compartments. Compartments were originally defined using the first eigenvector of Hi-C data and can be viewed as characteristic checker-board pattern in this data where regions from the same compartment show enriched interactions whilst less frequently interacting with regions found in the other compartment (Figure 1.1D) (Lieberman-Aiden et al., 2009). There are two main nuclear compartments that splits the genome into regions consisting of a few megabases that are reflective of their chromatin state: A and B. The A compartment mainly correlates with open, decondensed, accessible euchromatin and transcriptionally active and early replicating genes (Lieberman-Aiden et al., 2009). Conversely, the B compartment consists of closed, condensed and inaccessible heterochromatin with transcriptionally inactive and late replicating genes (Lieberman-Aiden et al., 2009). One other thing that has been found to highly correlate with the B compartment is regions that were identified by DamID of approximately 1 Mb in size that are found to be high contact with the nuclear lamina termed Lamina Associated Domains (LADs) (Guelen et al., 2008). Remarkably, these compartments correspond to the radial positioning of chromatin with the more repressed B compartment found at the nuclear periphery and around nucleoli while the more transcriptionally active A compartment found in the interior of the nucleus. The segregation of chromosomes into different compartments in individual nuclei has also been demonstrated using a multiplexed DNA FISH assay (Wang et al., 2016).

Subsequent work using an adapted *in situ* Hi-C protocol with increased resolution, it has been shown that A/B compartments can be more finely defined by their chromatin features into sub-compartments (Rao et al., 2014). The A1 sub-compartments corresponds to nuclear speckles and is typically more active than the A2 compartment (Rao et al., 2014). The B1 sub-compartment consists of facultative heterochromatin (e.g. Polycomb domains), pericentromeric heterochromatin is found in B2, B3 consists of LADs and the smallest sub-compartment B4 mainly contains small clusters of KRAB-ZNF genes found specific to human chromosome 19 (Rao et al., 2014). Again, this sub-compartmentalisation is suggestive that specific chromatin organisation into domains of similar chromatin states may be important for the different genomic functions of these compartments. Long-range clustering (>1 Mb) between polycomb-repressed genes has been observed both *in vivo* and *in vitro* and these polycomb domains are disrupted upon gene activation (Boyle et al., 2020).

The formation of these different membrane-less nuclear compartments has been proposed to be driven by a liquid-liquid phase separation (LLPS) mechanism (reviewed by e.g. Banani et al., 2017). Here, chromatin segregates according to mutual attraction with other chromatin of a similar ‘type’ and form droplets through multivalent interactions between chromatin binding proteins. This phenomenon is believed to be important for the formation of many structures within the nucleus such as the nucleolus, nuclear speckles and heterochromatin domains (Banani et al., 2017). There is also some evidence suggesting that phase separation has roles in gene regulatory processes through super enhancer clustering and the formation of transcriptional hubs or factories (discussed further in 1.2.2.3). Compartments have also been alternatively proposed to form by anchoring to specific regions such as the nuclear lamina and speckles (Chen et al., 2018b). Recent work using rod photoreceptors cells as a model to study compartmentalisation mechanisms, due to the inverted chromatin architecture observed in this cell type in nocturnal mammals, found that A/B compartmentalisation still occurs without heterochromatin tethering to the nuclear lamina arguing in favour of a phase separation mechanism (Falk et al., 2019). The authors also used polymer modelling to show that both conventional and inverted chromatin compartmentalisation was primarily driven by heterochromatin interactions.

HP1 has been shown to phase separate and is the likely driver of A/B compartmentalisation although more work is needed to fully elucidate if phase separation of heterochromatin is necessary for gene repression (Larson et al., 2017; Strom et al., 2017).

Both compartments strongly correlate with distinct functional processes - the active 'A' compartment vs the repressed 'B' compartment. Whether these different compartments form as a cause or consequence of their function is still under much investigation. Some work in favour of compartments influencing the chromatin state of a genomic region involved tethering a reporter gene to the nuclear lamina that resulted in gene repression (Reddy et al., 2008). A similar study using multiplexed random reporter integration to assay transcriptional output in different chromatin environments found a strong correlation between LADs and reporter gene repression (Akhtar et al., 2013). Another study using synthetic transcription factors to decompact a gene normally located in the B compartment resulted in its relocation to the nuclear interior (Therizols et al., 2014). Here, the authors targeted genes within a facultative LAD i.e. primed to move away from the periphery during differentiation. They showed that this gene movement was not sufficient to activate gene expression, decoupling transcription and chromatin compaction (Therizols et al., 2014).

1.1.1.3 Topologically associated domains and chromatin loops

At the sub-megabase scale, mammalian genomes are partitioned into TADs. Pioneering work using genome-wide interaction methods identified TADs and defined them as genomic regions showing enriched interactions between loci within their boundaries (Figure 1.1F) (Dixon et al., 2012; Nora et al., 2012). It is estimated that there are ~10,000 TADs in mammalian cells ranging from 40 kb to 3 Mb with a median size of 185 kb (Rao et al., 2014). Earlier work by the Cremer group had also identified ~1 Mb chromatin domains within the genome by imaging (Albiez et al., 2006). More recently, the presence of TADs as actual structures of chromatin within single cells rather than population-average features of Hi-C maps has been supported by super-resolution imaging of mega-base regions with Oligopaints and single-cell Hi-C experiments (Figure 1.1E) (Bintu et al., 2018; Flyamer et al., 2017).

Following the observation that TADs delineate intermediate structures of chromatin organisation, more features of TADs have become evident. TADs and their boundaries have been suggested to be conserved across cell types and different species (Dixon et al., 2012; Nora et al., 2012). This indicated that TADs may be important genome structures. TAD boundaries are typically enriched for housekeeping genes, tRNA genes, SINEs and CCCTC-binding factor (CTCF) binding sites (Dixon et al., 2012). A distinct high interaction peak can be seen in the corner of a number of TADs. The peak interaction is formed from two loci, often at large linear distances from one another, being brought into close proximity and forming a chromatin loop (Figure 1.1G) (Rao et al., 2014). Chromatin loops were found to be conserved across cell types and between mouse and human suggesting that loops may be important features of genome organisation in mammals (Rao et al., 2014). Chromatin loops can lead to transient interactions between distant loci or can be stabilised by architectural proteins forming more stable, invariant interaction. TADs have been proposed to represent multiple chromatin loops continually assembling and disassembling forming the genome-wide average heatmap depicted from e.g. Hi-C data (Fudenberg et al., 2016) (discussed further in 1.1.2). At some TAD borders near a cohesin loading site, a prominent stripe or track of high contact frequency can be seen (Kraft et al., 2019; Vian et al., 2018). This is indicative of a locus at a TAD boundary contacting across the entire TAD region at a higher frequency than other contacts. However, the function of these stripe features of TADs is unclear but it has been proposed that a promoter (or regulatory element) at the base of a stripe can track along the entire TAD and interact with other DNA elements (Kraft et al., 2019; Vian et al., 2018). Following increased sequencing depth, finer internal TAD structure has been observed. Nested within TADs, smaller TADs or sub-TADs can be seen (Hanssen et al., 2017). The functional significance of this intra-TAD structure is not well understood but there is emerging evidence that although TADs are largely invariant, intra-TAD structure can reconfigure between cell types (Dixon et al., 2012). These cell-type specific contacts are required for gene regulation during cellular differentiation and have been proposed to be mainly driven by transcription factors and chromatin modifications rather than genome architectural proteins (Andrey et al., 2017; Berlivet et al., 2013; Bonev et al., 2017).

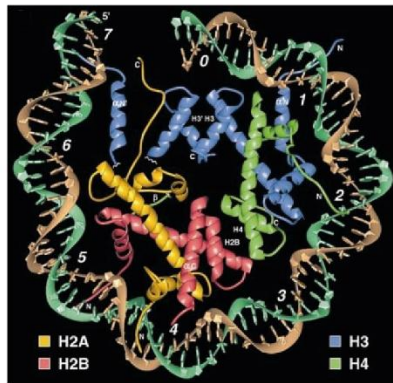
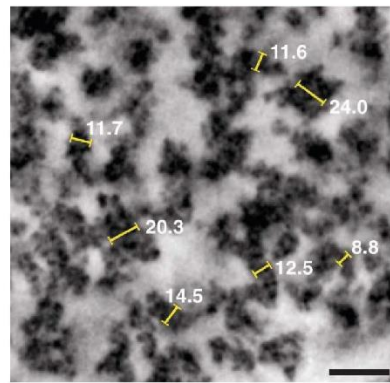
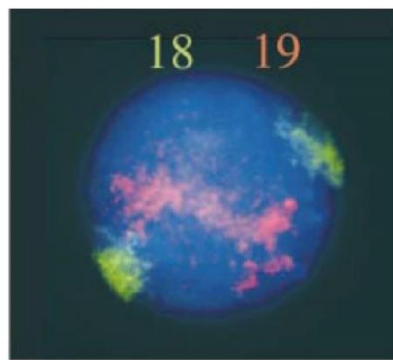
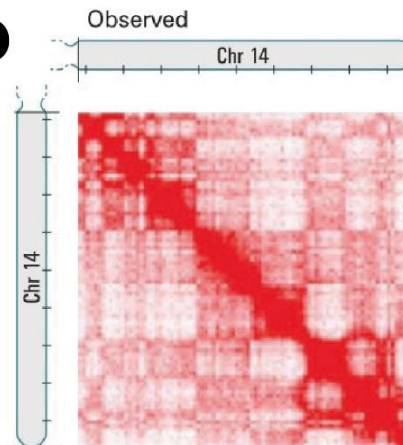
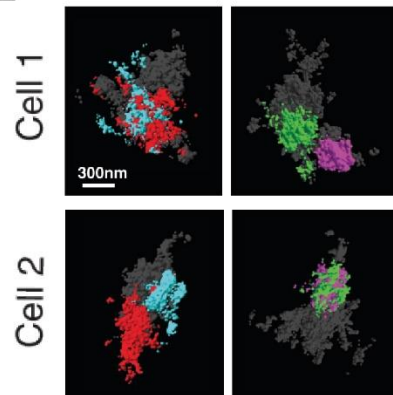
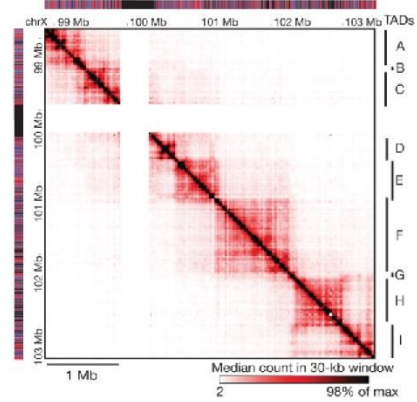
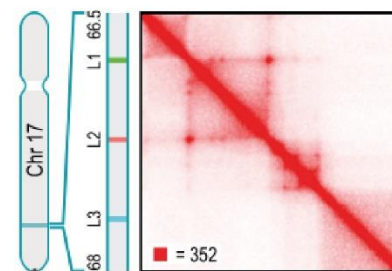
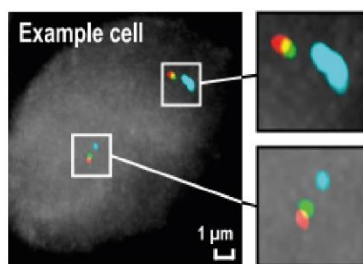
A**B****C****D****E****F****G**

Figure 1.1. Chromatin is organised in the nucleus across various scales.

(A) Crystal structure of the nucleosome core particle consisting of the four core histone proteins wrapped with DNA (taken from Luger et al., 1997). (B) ChromEMT images of chromatin that is folded into flexible, irregular structures with diameters ranging from 5-24 nm (taken from Ou et al., 2017). (C) Visualisation of chromosome territories using chromosome paints showing the gene-rich human chromosome 19 in the centre of the nucleus and the gene-poor chromosome 18 located closer to the nuclear periphery (taken from Croft et al., 1999). (D) Characteristic checker-board pattern observed for A/B compartments in Hi-C datasets for human chromosome 14 (taken from Lieberman-Aiden et al., 2009). (E) Representative super-resolution images of Topologically Associated Domain (TAD) structures showing their variability across single-cells using Oligopaints (adapted from Bintu et al., 2018). (F) 5C map of the mouse X-chromosome showing the TADs (denoted A-I) (taken from Nora et al., 2012). (G) 3D DNA FISH of a chromatin loop (left hand panel) on human chromosome 17 identified by Hi-C (right hand panel) (adapted from Rao et al., 2014).

1.1.2 Loop extrusion produces chromatin loops and TADs

Due to its high enrichment at mammalian TAD boundaries, CTCF is believed to be one of the main architectural proteins that is responsible for organising the genome into TADs. CTCF is an 11-zinc-finger, sequence-specific DNA-binding protein (Klenova et al., 1993; Lobanenko et al., 1990). It is highly conserved across bilaterian metazoans (Filippova et al., 1996; Heger et al., 2012) and found ubiquitously expressed across cell types with essential cellular and developmental functions (Moore et al., 2012; Soshnikova et al., 2010). CTCF was originally identified as an insulator or barrier protein in vertebrates (Bell et al., 1999). Here the authors demonstrated using a DNA fragment with known insulator activity from the chicken β -globin locus that this sequence was bound by CTCF and was required for enhancer blocking (Bell et al., 1999). CTCF has been further shown to be able to block enhancer function on reporter plasmids and within endogenous genomic locations (Bell and Felsenfeld, 2000; Burcin et al., 1997; Hark et al., 2000; Shrimali et al., 2012). CTCF is a multi-faceted protein with functions in gene activation/repression, genomic imprinting, X-chromosome inactivation, in addition to insulation (for review see e.g. Phillips & Corces, 2009) and chromatin looping (discussed further below).

It has been proposed that TADs, chromatin loops and stripes form through the binding of CTCF in an orientation-dependent manner at TAD boundaries and thereby,

constraining the extent of cohesin-mediated loop extrusion (Dixon et al., 2012; Fudenberg et al., 2016; Narendra et al., 2015; Nasmyth, 2001; Rao et al., 2014; Sanborn et al., 2015). This is supported by observations that the majority of CTCF sites are also occupied by cohesin (Parelho et al., 2008; Rao et al., 2014). However, it must be noted that not all TAD boundaries are bound by CTCF (Dixon et al., 2012). Cohesin is a ring-shaped protein complex (Figure 1.2A) which has a crucial role in sister chromatid cohesion and is known to topologically entrap DNA in an ATP-dependent manner (Murayama and Uhlmann, 2014, 2015). DNA entrapment by cohesin is not a requirement for loop extrusion in contrast to sister chromatid cohesion (Davidson et al., 2019; Kim et al., 2019). Cohesin-mediated loop extrusion has been recently demonstrated *in vitro* on DNA curtains (Davidson et al., 2019). In the loop extrusion model, a cohesin ring is loaded onto chromatin and tracks along whilst extruding DNA until it encounters a boundary element, proposed to be CTCF (Figure 1.2) (Fudenberg et al., 2016). This model is supported by the high correlation of convergent orientation of CTCF binding sites observed at TAD boundaries (Rao et al., 2014) and single-molecule imaging studies whereby the movement of cohesin along DNA is blocked by CTCF (Davidson et al., 2016; Labialle et al., 2014). Several studies have also tested the CTCF convergence rule using targeted manipulations of CTCF binding sites to show that convergent CTCF binding at distal sites limits cohesin extrusion and defines TAD boundaries (Guo et al., 2015; Narendra et al., 2015; Sanborn et al., 2015; de Wit et al., 2015). Crystal structure studies of CTCF/cohesin identified their binding interface, ‘conserved essential surface’ or CES (Li et al., 2020). This CES site is required for the formation and stabilisation of CTCF/cohesin-mediated chromatin loops (Li et al., 2020). The mechanism of loop extrusion provides cells with an active process to bring together distant loci whereas the long-range chromatin interactions formed through random collisions may happen too rarely to mediate biological functions.

The use of conventional genetic approaches such as genetic knockouts/knockdowns to study the functions of CTCF or cohesin have proven challenging due to the essential requirement of these proteins for cell division and embryonic development (Moore et al., 2012; Singh and Gerton, 2015; Soshnikova et al., 2010; Wan et al., 2008). Residual

protein can remain after protein knockdown and in conditional knockout mouse models leading to conflicting results in the literature (Figure 1.3A) (Seitan et al., 2013; Sofueva et al., 2013; Zuin et al., 2014). The development of the auxin-inducible degron (AID) system has overcome some of the issues associated with protein depletion experiments. In this system, target proteins tagged with an AID are ubiquitinated and degraded by the proteasome in the presence of auxin and TIR1 (Figure 1.3B) (Nishimura et al., 2009). The AID system provides a means to rapidly and reversibly deplete any target protein in mammalian cells.

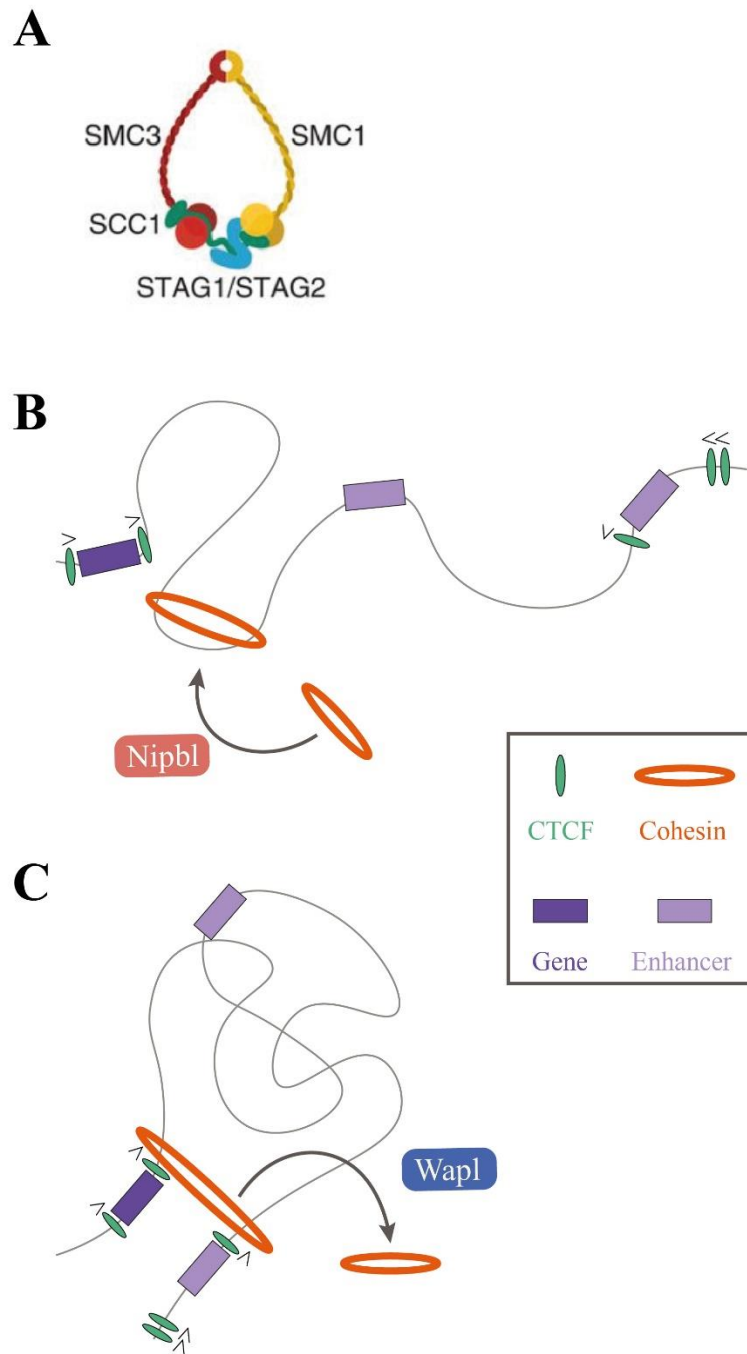


Figure 1.2. Loop extrusion produces topologically associated domains and chromatin loops.

(A) Diagram of the cohesin structure showing the different subunits (adapted from Davidson et al., 2019) (B-C) Simplified schematic of the loop extrusion process. (B) Cohesin is loaded onto chromatin by NIPBL. (C) Cohesin translocates along the chromatin fibre until it is blocked by CTCF bound in the convergent orientation (orientation denoted by arrowheads). Cohesin is removed from chromatin by WAPL.

Much of the evidence in favour of the loop extrusion model comes from work utilising the AID system or similar methods to deplete key molecular components of the model (summarised in Fudenberg et al., 2017). Removal of CTCF resulted in the diminishing of TAD boundaries, in support of CTCF defining TAD borders, whilst overall TAD structures remained and there was no effect on compartmentalisation (Figure 1.3C) (Nora et al., 2017; Wutz et al., 2017). In the absence of CTCF, cohesin is found to accumulate at active transcription sites, presumably halted by transcriptional complexes or other large bound proteins (Busslinger et al., 2017; Davidson et al., 2016). Cohesin depletion completely erased TADs and chromatin loops whilst compartments were strengthened (Figure 1.3D) (Rao et al., 2017; Wutz et al., 2017). The authors also showed that this removal of TADs was reversible and when cohesin was re-introduced into cells TAD structure was restored, demonstrating that cohesin is actively involved in TAD formation. Similar effects were observed when cells were depleted of the cohesin loader, NIPBL, suggesting that cohesin must be continually loaded onto chromatin in order to form TAD structures (Schwarzer et al., 2017). Although a small proportion of NIPBL binding sites overlap with CTCF, the vast majority do not suggesting that cohesin is loaded elsewhere on chromatin by NIPBL before extruding along (Figure 1.2B). Interestingly, following depletion of the cohesin unloader WAPL (Figure 1.2C) and its co-factors, larger TADs and novel loop peaks were observed and compartmentalisation was disrupted (Haarhuis et al., 2017; Wutz et al., 2017). The chromatin in these cells appeared very condensed with a characteristic ‘vermicelli’ appearance of chromosomes. This data demonstrates that the continual loading and unloading of cohesin is required to form TADs and chromatin loops and in order to also prevent chromatin becoming too condensed during interphase (Busslinger et al., 2017; Haarhuis et al., 2017; Liu et al., 2020; Wutz et al., 2017). This observation of the dynamic nature of loop formation is supported by single-molecule tracking of CTCF and cohesin molecules which found that CTCF and cohesin have relatively short residence times on chromatin (Hansen et al., 2017). Another study which took advantage of the AID system to deplete cohesin genome-wide constructed distance matrices from super-resolution imaging of mega-base domains of chromatin from thousands of cells (Bintu et al., 2018). The author’s found,

consistent with previous work, TADs were eliminated at the population-wide level following cohesin degradation. However, upon examination of single-cells, they found TAD-like structures persisted although the boundaries of these domains varied across cells rather than the typical conserved TAD boundary position (Bintu et al., 2018).

How the process of extrusion by cohesin is driven along chromatin has also been investigated. It has been proposed that the extrusion is powered by (1) supercoiling of DNA as a result of transcription (2) thermal diffusion or (3) the ATP-dependent activity of cohesin. Studies have shown that cohesin is translocated along chromatin by active transcription (Busslinger et al., 2017; Davidson et al., 2016; Racko et al., 2018). Cells that have been depleted of ATP show an accumulation of cohesin molecules at NIPBL binding sites and they cannot move cohesin along DNA as effectively when the ATPase domain of cohesin is mutated (Vian et al., 2018). This data supports that extrusion requires the ATPase-activity of cohesin (Vian et al., 2018). This work is consistent with another study which proposed that NIPBL continually associates with cohesin on chromatin to stimulate its ATPase domain and the extrusion process (Rhodes et al., 2017). Contradictory to previous reports, inhibition of transcription or replication had no effect of the formation of chromatin loops (Vian et al., 2018). Another member of the SMC family with a similar ring structure to cohesin; condensin, was recently demonstrated by live-cell imaging to extrude DNA in an ATP-dependent manner (Ganji et al., 2018).

Together, these perturbation studies together with polymer modelling strongly support loop extrusion as the mechanism of TAD formation with cohesin and its related proteins involved in forming chromatin loops and CTCF defining TAD boundaries within mammalian cells. This work also indicates that higher-order chromatin organisation is not formed in a hierarchical manner. Rather, compartments and TADs form through independent mechanisms and loop extrusion may even function to antagonise compartmentalisation (Nuebler et al., 2018; Pereira et al., 2018; Schwarzer et al., 2017).

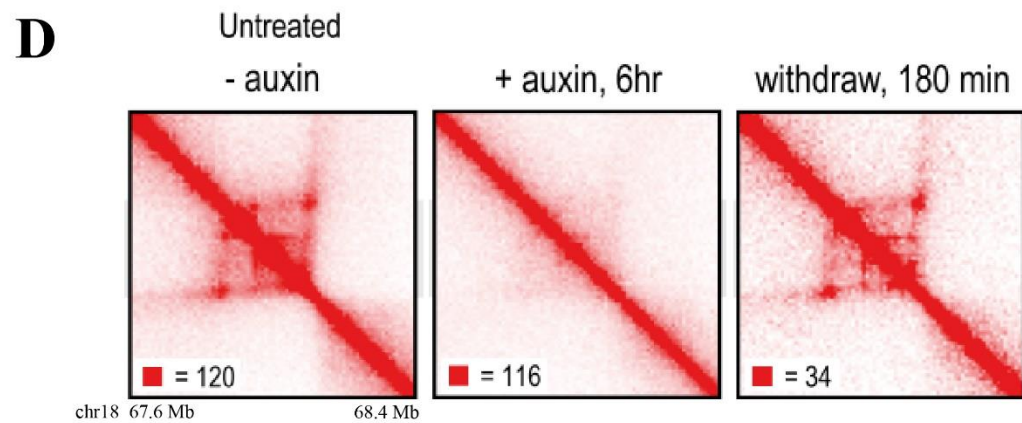
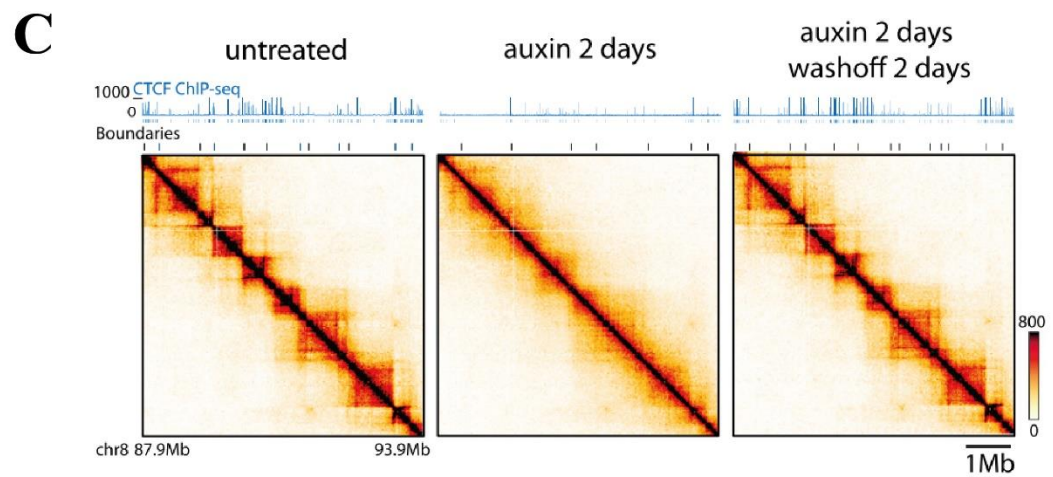
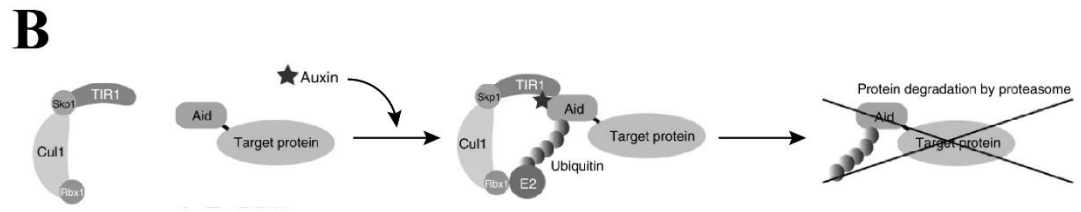
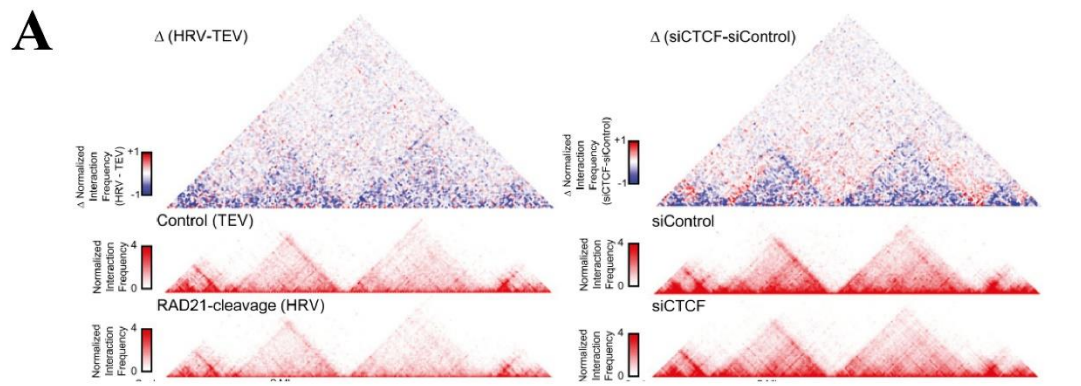


Figure 1.3 Depletion of CTCF or cohesin as a means to study TAD functions.

(A) TAD structure persists after conditional knockout of cohesin (left-hand graphs) or siRNA-mediated knockdown of CTCF (right-hand graphs) (adapted from Zuin et al., 2014). (B) Schematic of the AID system. In the presence of auxin, aid-tagged proteins will interact with TIR1 leading to ubiquitination of the tagged protein and subsequent degradation by the proteasome (adapted from Nishimura et al., 2009). (C-D) The AID system can be used to reversibly deplete (C) CTCF or (D) cohesin genome-wide causing loss of chromatin loops and TAD structure disruption (adapted from Nora et al., 2017; Rao et al., 2017).

1.1.3 TADs as a functional unit of gene regulation

The regulatory landscapes of several key developmental genes have frequently been found to overlap with these TAD structures defined by chromosome conformation capture based methods (Andrey et al., 2017). This suggests that regulatory elements will generally lie within the same TAD as their target promoter. As TADs are defined as genomic regions which show enriched interactions between loci within their boundaries (see 1.1.1.3) this has led to speculation that the structuring of chromatin into TADs serves as a means of reducing the 3D search space of regulatory elements for their cognate gene. The observation of the strong association between domain boundaries and the binding of the insulator protein CTCF also indicated that there was a link between TAD structure and gene regulation as CTCF was known to act as a barrier to heterochromatin spreading, to block enhancer activity and to also mediate long-range regulatory interactions (reviewed by e.g. Ong and Corces, 2014).

Below, I discuss some of the work investigating TADs as important gene regulatory domains and the links between TAD perturbation and human disease.

1.1.3.1 Correlation between TADs and regulatory landscapes

One study utilised a regulatory sensor that was integrated into chromatin and allowed the authors to map the regulatory activities of a reporter gene at different positions across the genome (Symmons et al., 2014). The regulatory domains identified using this reporter overlapped substantially with TADs and were limited by TAD boundaries. This genome-wide correlation between regulatory landscapes and TADs supported the notion that TADs function both in chromatin organisation and gene regulation.

Another study using mouse embryonic limb bud tissue from various stages of development combined Capture-C and ChIP-seq to define stable regulatory interactions that were marked by CTCF/cohesin and most likely represent the underlying invariant TAD structure (Andrey et al., 2017).

Genes within a TAD can be co-regulated suggesting that this co-expression may be important during cell differentiation and development (Andrey et al., 2013; Nora et al., 2012). By careful dissection of both the active and inactive X-chromosome, the authors showed that two key non-coding RNAs involved in X inactivation, *Xist* and *Tsix*, were located in distinct TADs with a strong intervening boundary. Interestingly, genetic elements known to be involved in *Xist* expression and thus *Xists*' regulatory landscape were contained inside the same TAD (Nora et al., 2012). Removal of the boundary between the *Xist* and *Tsix* TADs led to increased interactions between then neighbouring TADs and gene misregulation (Nora et al., 2012). This work exemplifies that TADs are important determinants of regulatory landscapes by preventing cross-talk across its boundaries and ensuring that co-regulated genes are constrained together. Also, gene co-regulation within a spatial "hub" would allow the sharing of important resources such as RNA Pol II clusters and/or regulatory elements.

A number of studies have also used genome engineering approaches to disrupt chromatin loops and/or TAD boundaries by deleting or inverting CTCF sites. As well as demonstrating the CTCF convergence rule (see 1.1.2), these experiments provided functional evidence that TADs normally facilitate regulatory landscape interactions and manipulation to these CTCF binding sites can promote TAD boundary crosstalk (Downen et al., 2014; Guo et al., 2015; Narendra et al., 2015; Rodríguez-Carballo et al., 2017; Sanborn et al., 2015; Williamson et al., 2019; de Wit et al., 2015).

Whilst there is compelling work indicating that TADs represent important units of gene regulation, when components of TADs are depleted and TADs disrupted at the genome-wide level the picture is less clear (Figure 1.3). Contradictory reports using the auxin-inducible degron system to acutely deplete CTCF or cohesin have shown that loss of either of these components has limited effects of gene expression despite the weakening of TAD boundaries (CTCF removal) or elimination of TADs (cohesin

removal) (Haarhuis et al., 2017; Kubo et al., 2017; Nora et al., 2017; Rao et al., 2017). Likewise, limb-specific conditional knock-outs of CTCF resulted in relatively modest gene expression changes (Soshnikova et al., 2010). Conversely, *in vivo* studies using a conditional knock out of the cohesin loading factor, *Nipbl*, demonstrated that TADs and loops were lost which led to some gene expression changes that were postulated to be due to ectopic enhancer activation in the absence of TAD insulation (Schwarzer et al., 2017).

It remains unclear precisely how or whether TADs function during mammalian development to maintain robust gene expression domains. A recent study used the highly rearranged balancer chromosomes from *Drosophila* to address whether chromosome-wide changes to TAD organisation and chromatin loops had consequences to gene expression profiles (Ghavi-Helm et al., 2019). Remarkably only modest gene expression changes were observed despite the extensive chromosomal rearrangements suggesting that regulatory interactions were preserved (Ghavi-Helm et al., 2019). It must be noted that balancer chromosomes are under selection so may not be truly representative of regular gene regulatory processes.

1.1.3.2 TAD disruption in human disease

Enhancer adoption or hijacking occurs when structural variants disrupt chromatin structure causing ectopic enhancer activity and subsequently human disease (Lettice et al., 2011; Northcott et al., 2014). There have been a number of studies reporting enhancer adoption as the underlying disease mechanism for various different human developmental disorders.

Lupiáñez et al described three limb abnormalities that were caused by different structural variants that overlapped with the *EPHA4* TAD (Figure 1.4). By carefully engineering these mutations into mouse models using CRISPR-Cas9, the authors demonstrated using 4C-seq that ectopic interactions could be detected between the *Epha4* TAD and genes found in the neighbouring TAD in both patient-derived cells and mutant mouse limb tissue (Figure 1.4). They found that upon disruption to TAD boundaries, ectopic interactions occurred between regulatory elements in the *EPHA4* TAD and non-target genes located in adjacent TADs (Lupiáñez et al., 2015). It is

believed that the ectopic gene activation through enhancer adoption, as shown in the mutant mouse models during embryonic development, is what causes the limb defects observed in the patients (Figure 1.4) (Lupiáñez et al., 2015). Similar studies have also described other chromosomal rearrangements identified in patients with human pathologies that disrupt gene regulation by fusing or shuffling TAD structures (Franke et al., 2016; Kraft et al., 2019).

The enhancer adoption phenomenon due to mutated TAD boundaries have also been observed in human cancers. One notable study showed that gain-of-function mutations in the *IDH* gene led to increased DNA methylation genome-wide in glioma cell lines (Flavahan et al., 2016). This gain of methylation reduced CTCF binding at some TAD boundaries and in one example, disruption to the boundary led to the oncogenic activation of *PDGFRA* through ectopic interactions with an off-target regulatory elements (Flavahan et al., 2016). Similar work has also shown that TAD boundary deletions can induce oncogene activation in several different cancers (Hnisz et al., 2016).

Together these studies show that altered TAD structure can have pronounced effects of gene expression that ultimately cause human disease. One thing that remains unclear is whether chromatin is structured into TADs as a result of gene expression or whether regulatory landscapes are defined by TADs. Also, whether only a subset of genes require TADs to constrain their regulation is uncertain.

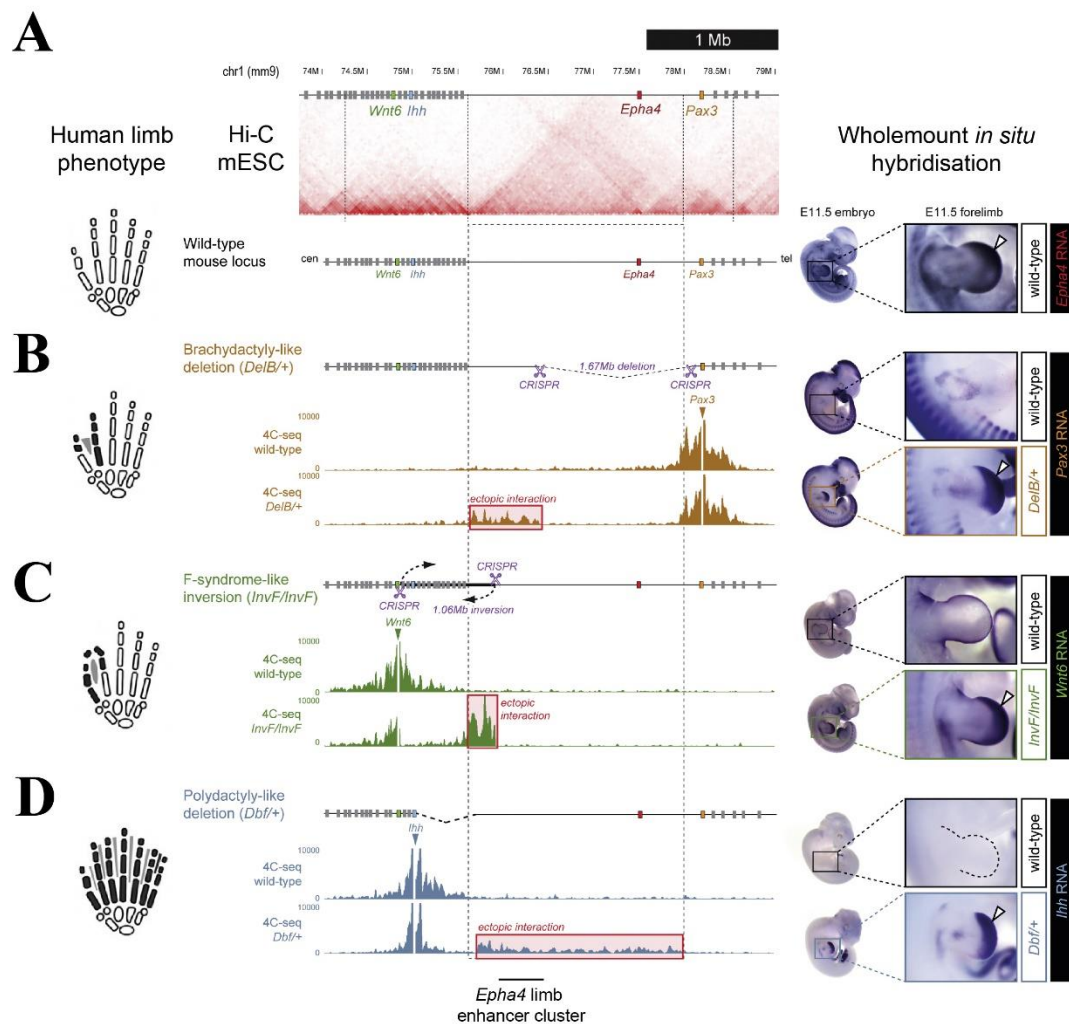


Figure 1.4. Human structural variants disrupt TADs and can have pathological consequences due to enhancer adoption.

Summary of the different pathological structural variants discussed in Lupiáñez et al., (2015). (A) Schematic of the *Wnt6/Ihh/Epha4/Pax3* TAD structures using Hi-C data from wild type mESCs with a map of the regulatory landscapes below. A normal human limb is shown on the left and on the right shows the normal limb expression pattern of *Epha4* in an E11.5 embryo. Dashed lines indicate TAD boundaries. (B) Brachydactyly patients display shorted digits and webbing. Brachydactyly-like deletion in mouse removes a TAD boundary leading to ectopic expression of *Pax3* in the limb due to altered interactions between *Pax3* and the *Epha4* limb enhancers. (C) F-syndrome patients display digit syndactyly. F-syndrome-like inversion in mouse moves *Epha4* limb enhancers into the *Wnt6* TAD leading to ectopic interactions and mis-expression on *Wnt6* in limbs. (D) Polydactyly patients have extra digits. Polydactyly-like deletion in mouse removes a TAD boundary inducing *Ihh* expression in the limb due to ectopic interactions with *Epha4* limb enhancers. This figure was adapted from Lupiáñez et al., (2015).

1.2 Gene regulation during development

Complex organisms are composed of multiple cell and tissue types that arise from single-cell embryos. The generation of multiple different cell types during mammalian development is reliant on *cis*-regulatory elements. The main functional classes of *cis*-regulatory elements are promoters, enhancers, repressors and insulators. Enhancers are essential for defining proper spatiotemporal gene expression patterns during embryonic development.

In the sections below, I will define what an enhancer is, what features are characteristic of these elements and how enhancers are believed to control gene expression, particularly with a focus on how long-range enhancers operate. I will also detail how disruption to enhancer function can cause human disease.

1.2.1 An overview of developmental enhancers

1.2.1.1 Enhancers have distinct chromatin features

Enhancers were originally defined as non-coding, *cis*-acting DNA sequence elements that were capable of ‘enhancing’ transcription of an *in vitro* reporter gene in a position- and orientation-independent manner (Banerji et al., 1981; Moreau et al., 1981). These regulatory elements are now known to define cell type-specific gene expression profiles (for review see e.g. Heinz et al., 2015). It is estimated that mammalian genomes contain between hundreds of thousands to millions of enhancers (ENCODE Project Consortium, 2012). An interesting feature of enhancers that emerged, particularly with regards to developmental genes, is that they are evolutionarily conserved with some being described as ultra-conserved due to 100% sequence similarity between humans and rodents (Bejerano et al., 2004). Comparative genomics is an extremely useful approach to identify non-coding DNA sequences with potential to harbour regulatory potential for key developmental loci (Nobrega et al., 2003; Woolfe et al., 2005). However, not all regulatory elements are conserved between species. Species-specific regulatory sequences have been proposed to confer phenotypic variation that are important for divergent nuclear hormone regulated transcriptional response in mammals (Jubb et al., 2016) and to drive morphological change important for adaptive evolution in stickleback fish (Jones et al., 2012).

Enhancers act as platforms for the binding of cell type-specific transcription factors (Long et al., 2016). Enhancer activation involves the binding of pioneer factors that can recruit nucleosome remodelling complexes that form a nucleosome-free region of chromatin and enzymes to chemically modify the surrounding histone proteins (Long et al., 2016). Poised or primed enhancers are typically marked by H3K4me1/2 and the absence of acetylation (He et al., 2010; Heintzman et al., 2007, 2009). Following binding of cell type-specific transcription factors, further chromatin opening, acetylation of H3K27 and H3K122 as well as recruitment of Mediator are all involved in enhancer activation (Creyghton et al., 2010; Heintzman et al., 2009; Pradeepa et al., 2016; Rada-Iglesias et al., 2011). Enhancer activation can be accompanied by transcription of enhancer RNAs. These shared biochemical features of enhancers has led to their genome-wide annotation across different cell and tissue types, through projects such as the Encyclopaedia of DNA Elements (ENCODE) Consortium (ENCODE Project Consortium, 2012), and is currently one of the main approaches to identify DNA sequences that contain candidate enhancers. As discussed above, not all enhancers are conserved between species and alternate approaches such as p300 chromatin immunoprecipitation (ChIP) have been found as a better predictor of enhancer activity rather than comparative genomics and has the advantage of identifying the cells where the enhancer of interest functions (Blow et al., 2010; Visel et al., 2009). The development of high-throughput methods such as DNase-seq and ATAC-seq provide a way in which to scan genome-wide for enhancers using DNase hypersensitivity or open chromatin as a proxy for enhancers (Boyle et al., 2008; Buenrostro et al., 2013). The original enhancer definition still holds today but has been expanded as follows; a DNA sequence that must be shown to increase transcription of its target gene in a *cis*-acting manner within its endogenous (*in vivo*) context when bound by cell type-specific TFs and with the associated chromatin/biochemical features.

1.2.1.2 Enhancers can behave additively or redundantly

Many genes have multiple enhancers with each element having precise tissue-specific activities that can function additively or redundantly during development. Estimates suggest that on average every human gene has ~10-20 enhancers but this is based on

enhancers identified solely by their biochemical features and have not been functionally validated (ENCODE Project Consortium, 2012). Enhancers that function in an additive manner generate a complex gene expression pattern by each having distinct regulatory activities in different tissue for same target gene (Long et al., 2016). This means enhancer loss only impacts its target tissue rather than the entire gene expression profile. Additionally, additive enhancers are believed to have been exploited during morphological evolution to acquire novel functions e.g. limb development (for review see Petit et al., 2017). Alternatively, enhancers can function redundantly. Redundant (also termed shadow or secondary) enhancers show highly overlapping gene regulatory activities meaning loss of one enhancer can be compensated by its redundant partner (Osterwalder et al., 2018). In one study when redundant enhancers were deleted on a sensitised background, developmental defects were observed suggesting redundancy between these elements normally provides regulatory robustness (Osterwalder et al., 2018). Some enhancers have been further classified as super-enhancers. These elements are defined as clusters of enhancers highly enriched for transcriptional co-activators such as Mediator and typically regulate key cell type identity genes e.g. pluripotency genes in embryonic stem cells (Hnisz et al., 2013; Whyte et al., 2013).

1.2.2 Mechanisms of long-range enhancer regulation

One of the most remarkable features of enhancers is that they can be found at large distances away in the genome from their target promoter with examples of distal enhancers over 1 Mb away (Long et al., 2020). Developmental genes are often surrounded by large regions of DNA devoid of any other genes termed ‘gene deserts’ (Nobrega et al., 2003). Long-range or distal enhancers are often found scattered throughout these gene deserts (Ovcharenko et al., 2005). Localised hopping activity of a sleeping beauty transposon was used to probe regulatory landscapes such as *Shh* and showed that enhancers can behave promiscuously across large genomic distances (Anderson et al., 2014; Symmons et al., 2014, 2016). Enhancers may confer some promoter specificity and they do not necessarily regulate every gene within a given regulatory landscape or TAD (Anderson et al., 2014; Symmons et al., 2016). Why some enhancers are found at large distances from their target gene and how they

enhance transcription from afar is not well understood. It should be noted that long-range enhancers do not require large genomic distances for their function. Long-range enhancers can act upon reporter constructs integrated close by (Anderson et al., 2014; Symmons et al., 2016) and large genomic deletions that bring enhancers into close genomic distance to their target promoter do not disrupt enhancer function (Williamson et al., 2019). One potential explanation for the presence of these long-range enhancers is that mammalian genes are estimated to be regulated by multiple enhancers meaning large distances allows the enhancers to be spread out across gene deserts; therefore, preventing their interference with each other. Alternatively, enhancers operate within TADs so although the linear distance may be large the spatial distance can be significantly reduced by chromatin folding meaning over time enhancers with close spatial proximity may have been selected for. A number of mechanisms have been proposed to explain how these enhancers achieve long-range gene regulation (reviewed by e.g. Benabdallah and Bickmore, 2016). These models of enhancer-promoter communication include; tracking, linking, looping and phase condensates (reviewed by e.g. Furlong and Levine, 2018).

1.2.2.1 Tracking and linking models

In the tracking model of enhancer regulation, the transcriptional machinery is recruited to the enhancer and transcribes along the intervening chromatin toward the target promoter (Figure 1.5A) (Furlong and Levine, 2018). Alternatively, the linking model involves recruitment of transcription factors to the enhancer that then oligomerise forming a protein bridge that links together the enhancer and promoter (Figure 1.5B) (Furlong and Levine, 2018). Both of these models involve the spread of a signal that originates at the enhancer and terminates at the promoter. However, a key difference is that at the end of tracking along the chromatin, the enhancer is brought along with the transcriptional machinery forming a loop with the promoter whereas in the linking model the enhancer and promoter remain distant from each other.

1.2.2.2 Looping model

Another model of long-range enhancer-promoter communication is the looping model where a distal enhancer is looped into close proximity to its target promoter allowing

direct contact between the two DNA elements (Figure 1.5C) (Furlong and Levine, 2018). In this model the direct contact between the enhancer and promoter would allow regulatory signals/information to be transferred. Chromatin loops are often found to contain promoters and enhancers at their base and are associated with gene activation (Rao et al., 2014). As discussed in 1.1.3.1, the regulatory landscape of genes are typically contained within TADs, suggesting that TADs function to constrain enhancer activity, and TADs have been shown to form through loop extrusion. Interestingly, extrusion would provide a mechanism by which distal enhancers loop to interact with their target promoters. The dynamic nature of loop formation could facilitate tissue-specific enhancers finding their cognate promoter as cells differentiate. Enhancer-promoter looping was also shown to be cell-type specific which agrees with the notion that enhancers have the ability to activate gene expression in a tissue-specific manner during cell differentiation (Tolhuis et al., 2002). In addition to CTCF/cohesin, other proteins have also been implicated in forming enhancer-promoter loops including; YY1, LDB1 and various transcription factors (Bonev et al., 2017; Krivega and Dean, 2017; Schoenfelder et al., 2010; Weintraub et al., 2017).

One of the most compelling demonstrations of looping as a means of gene regulation involved engineering of a loop between the *β -globin* promoter and the LCR. The authors showed that this loop was sufficient to activate expression of *β -globin* in cells that do not normally express this gene thereby directly implicating enhancer-promoter looping as a mechanism to activate gene expression over a distance (Deng et al., 2014). More recently, live-cell imaging of *Drosophila* embryos showed that transcriptional bursting was correlated with looping interactions between a distal enhancer and promoter (Chen et al., 2018a). A subsequent live-cell imaging approach of the *Sox2* locus in mammalian embryonic stem cells found contradictory results with no correlation between enhancer-promoter proximity and transcription (Alexander et al., 2019). Further analysis is required to resolve these discrepancies – the species used, gene of interest and genomic labelling methods differ across the studies. Live-cell imaging of other mammalian genes in the future should hopefully reconcile these studies. In addition, depletion of CTCF/cohesin, key proteins shown to form DNA loops, has limited effects on gene expression despite genome-wide disruption to

chromatin loops (see 1.1.3.1) (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017). The requirement of direct enhancer-promoter interaction being necessary and sufficient for gene activation is still under much debate.

1.2.2.3 Phase condensate model

As discussed in 1.1.1.2, LLPS has been proposed to be the mechanism by which chromatin organises into compartments, mainly driven by the association between heterochromatin. An LLPS mechanism has also been implicated in gene regulation. In this model, proteins involved in transcriptional regulation interact together through weak hydrophobic interactions between their intrinsically disordered regions (IDRs) (Banani et al., 2017; Hnisz et al., 2017). These multivalent interactions create a high concentration of proteins in a phase condensate that facilitate biochemical reactions involved in gene activation (Figure 1.5D). A hub containing a high concentration of proteins important for transcription was also described for the previously proposed ‘transcriptional factories’ (for review see e.g. Buckley and Lis, 2014). However, in the transcription factory model these ‘hubs’ pre-exist and genes move into them to be activated. It has been reported that the LCR of the β -globin locus is found to associate with transcriptional factories (Noordermeer et al., 2008). Whereas in the phase condensate model the ‘hub’ form on chromatin.

A trio of studies were published back to back in 2018 that investigated the LLPS behaviour of proteins involved in transcriptional control. All of the studies used imaging based approaches to show that transcriptional activation proteins including Mediator, p300 and RNA polymerase II form puncta inside the nucleus that are sensitive to 1,6-hexanediol treatment (Cho et al., 2018; Chong et al., 2018; Sabari et al., 2018). 1,6-hexanediol is frequently used to investigate LLPS properties of proteins as it disrupts hydrophobic interactions between IDRs and effectively dissolves any phase condensates. One of these studies focused on the LLPS behaviour of proteins found at high concentrations at super-enhancers; Mediator and BRD4 (Sabari et al., 2018). Here, recruitment of transcription factors and coactivators at super-enhancers leads to the formation of a high concentration condensate that allows activation of the target gene that does not depend on direct interaction between the enhancer and

promoter. Subsequent work demonstrated that both transcription factors and oestrogen receptors can form condensates with Mediator (Boija et al., 2018). The observation that different proteins involved in transcriptional activation can form phase condensates suggests that this model may be a more general mechanism that allows a diverse set of trans-activating proteins bound at different enhancers to interact with a core set of co-activators leading to gene expression. In contrast to the looping model, multiple long-range enhancers could interact with a transcriptional condensate and thereby cooperatively activate their target promoter without the need for each enhancer to directly interact (Alexander et al., 2019; Benabdallah et al., 2019; Lim et al., 2018). In accordance with this, it was shown that a shared enhancer can induce transcriptional bursting of two genes simultaneously, challenging the looping model as it would be expected if an enhancer and promoter looped together then each gene would show distinct transcriptional dynamics (Fukaya et al., 2016). The coordinated gene regulatory action of this enhancer is more consistent with the phase condensate model.

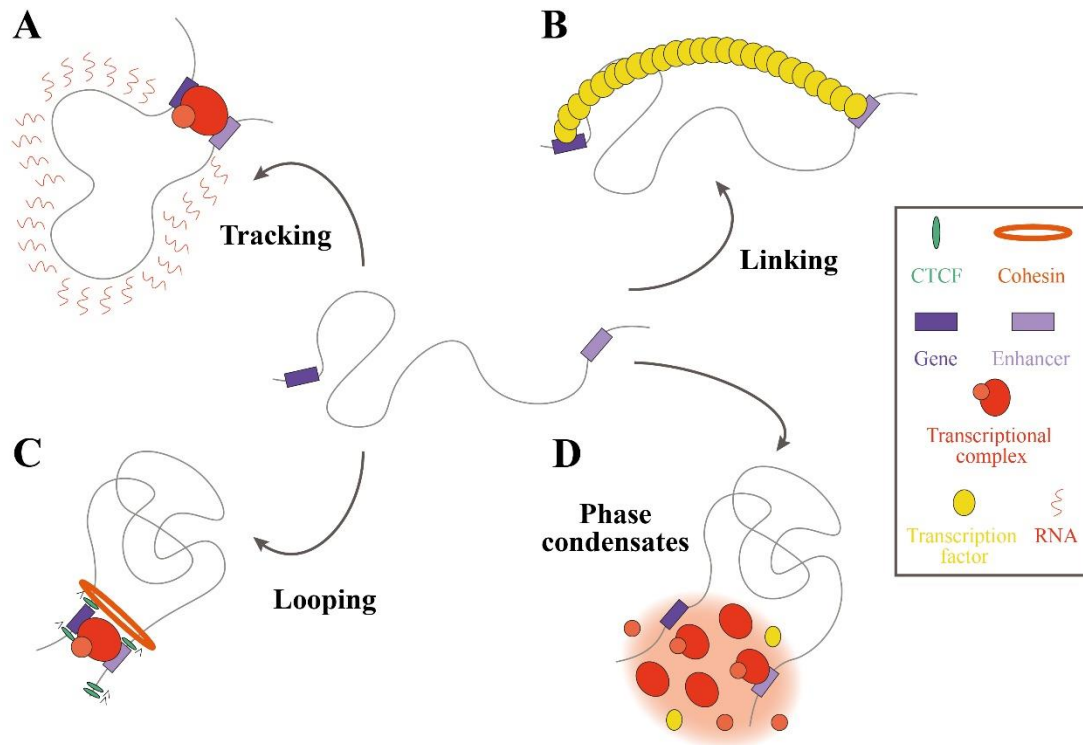


Figure 1.5. There are a number of models of long-range enhancer regulation.

Simplified representation of long-range enhancer regulation models. **(A)** Tracking involves transcriptional machinery recruitment to the enhancer that then transcribes along intervening chromatin to the target promoter. **(B)** Linking involves recruitment of transcription factors to the enhancer that oligomerise into a protein bridge linking together the enhancer and promoter. **(C)** Looping brings a distal enhancer into close proximity to its target promoter allowing direct interaction between the two DNA elements. Here this looping is mediated by CTCF/cohesin but there are other candidate looping factors. **(D)** Phase condensates form from recruitment of transcription factors and coactivators at enhancers creating a high concentration of transcriptional machinery leading to gene activation.

1.2.2.4 Comparison of models

Common to all the models of long-range enhancer-promoter communication is a reliance on DNA binding proteins (CTCF, transcription factors) and transcriptional machinery (RNA pol II, mediator, cohesin). C-based and imaging studies show that some long-range enhancer-promoter interactions are invariant and appear to be pre-established before enhancer activation (reviewed by e.g. Andrey and Mundlos, 2017). These invariant contacts have been proposed to be maintained by CTCF/cohesin

chromatin or alternatively, by the Polycomb Repressive Complexes. Pre-formed contacts would facilitate rapid induction of gene expression during development. However, stable enhancer-promoter contacts are not compatible with genes that have many enhancers with distinct spatiotemporal activities. Indeed, some developmental genes form dynamic regulatory interactions where distal enhancer-promoter contacts appear concurrently with enhancer activation and target gene expression (Andrey and Mundlos, 2017). For example, the LCR of the globin locus shows differential contacts with the embryonic and adult globin genes at different stages of erythroid development (Palstra et al., 2003). A major question is whether enhancers must have a sustained interaction with their target promoter to maintain expression or whether a transient interaction is sufficient for gene induction. With the exception of the linking model, the various different models could be applied to either of these scenarios.

The different models of long-range enhancer-promoter communication have some overlapping features and these proposed models or at least some of their features could potentially co-exist to fine-tune mammalian gene regulation. Loop extrusion and LLPS are independent mechanisms driving chromatin organisation (see 1.1.2) and I speculate extrusion may be involved in facilitating tissue-specific enhancers in their search for their cognate promoter over large distances whilst gene activation may involve both of these elements becoming associated with a transcriptionally active phase condensate. The number of chromatin loops detected by Hi-C are less numerous than the number of predicted enhancers in the mammalian genome (Rao et al., 2014). Generally, TADs are invariant across tissues, although intra-TAD structure can reconfigure between cell types possibly reflecting tissue-specific regulatory interactions (see 1.1.1.3). Perhaps, the extrusion process is sufficient to allow the enhancer to closely interact with its target promoter and initiate transcription but does not necessarily need to maintain a close contact. The CTCF at the TAD boundaries are there to keep the interactions within the TAD by limiting the extrusion process and insulate activities between enhancers and non-target genes.

As discussed in 1.2.1, enhancers can sometimes bypass other genes in order to regulate expression of their target. It is difficult to envision how the tracking or phase

condensate models may exclude intervening genes from being subject to activation. Long-range regulation via linking or tracking may also be incompatible with enhancers found up to a 1-2 Mb away. Cohesin has been shown to translocate or track rapidly along chromatin (Davidson et al., 2016). “One-size-fits-all” with regards to mechanisms of long-range gene regulation may not be appropriate for every enhancer-gene pair. Also different mechanisms may apply to stimulus-response genes versus developmental loci.

1.2.3 Enhancer dysregulation in human disease

Human genomes contain around 20,000 genes and active genes have been estimated to be regulated by multiple different regulatory elements in a cell type-specific manner (Sanyal et al., 2012). Therefore, genetic variation or mutation in enhancers should be associated with human disease. Genome-wide association studies (GWAS) have revealed that the majority of disease- and quantitative trait- associated common genetic variation (SNPs) is found in the non-coding region of the genome with many of these GWAS hits found within distal enhancers (Noonan and McCallion, 2010). There is a great interest on characterising how enhancers normally function during mammalian development so we can understand how disruption to enhancers can cause human disease.

A number of human diseases are due to rare/*de novo* mutations in enhancers. One of the classic examples of an enhancer mutation underpinning a human genetic disorders is disruption to the LCR of the β -globin gene causing β -thalassemia (Kioussis et al., 1983). Leri-Weill syndrome, known to be caused by mutations in the *SHOX* gene resulting in short stature in patients, has been demonstrated to be phenocopied in patients with *SHOX* enhancer deletions (Chen et al., 2009). Deletions that encompass extreme long-range enhancers (1.45 Mb) has also been found in patients with Pierre Robin Sequence (Long et al., 2020). Also, both gain- and loss-of-function mutations have been identified in the enhancers of *Shh* resulting in distinct human developmental disorders (discussed further in 1.3.6). Another disease mechanism related to enhancers involves an increase in enhancer copy number causing an increase and misexpression of its target gene. This was shown for the *Ihh* locus where duplications of its redundant

enhancers upregulated *Ihh* expression leading to developmental defects such as syndactyly, polydactyly and craniosynostosis in patients (Will et al., 2017).

One of the major aims of developmental biology is to understand how our genes function and how perturbations to gene function contribute to developmental defects. These insights provide patients, family members and clinicians with a better understanding of these sometimes debilitating genetic disorders and highlight areas that may be targeted in the future for therapeutic intervention. It is now apparent that not only mutations that directly affect enhancer activity cause human disease but as discussed in 1.1.3.2, structural variants that modify TADs can directly contribute to a number of severe developmental disorders by disrupting or redirecting regulatory interactions. This makes identification of the affected gene in patients with mutations in noncoding regions far more complex as chromatin conformation must also be considered. Compared to coding mutations, aberrations of enhancer activity or function can have more subtle effects on gene expression and may only affect specific tissues or developmental stages due to the strict spatio-temporal activities of enhancers. These complexities highlight the need for *in vivo* methods for both enhancer characterisation and evaluation of mutations. One recent study described a zebrafish reporter assay that has been developed for rapid quantitative assessment of enhancer mutations using a dual fluorescent reporter that allows the affected cell/tissue types and relevant developmental time windows to be identified (Bhatia et al., 2020).

1.3 *Shh*: A model locus of gene regulation

Sonic Hedgehog (Shh) is a developmentally regulated morphogen encoded by the *Shh* gene that has critical growth and patterning functions during embryonic development. The *Shh* regulatory domain has emerged as a paradigm locus for the study of long-range enhancer regulation.

In this section, I will detail the function of the *Shh* gene, its cellular signaling pathway, how its expression is regulated, what is known about the structure of the *Shh* TAD and how dysregulation to *Shh* causes developmental disorders.

1.3.1 *Shh* in embryonic development

1.3.1.1 Identification of the *hedgehog* genes

A seminal study carried out by Nüsslein-Volhard and Wieschaus (1980) involved a large-scale systematic mutagenesis screen of *Drosophila* to identify genetic mutations that affect segmental patterning. This led to the original identification of the *hedgehog* gene whose mutation led to abnormal segmental polarity (Nüsslein-Volhard and Wieschaus, 1980). *Hedgehog* mutations are embryonic lethal and mutants display small pointy projections, called denticles, across the entire ventral side of the *Drosophila* embryo which had a resemblance to the spikes of a hedgehog, hence the given gene name (Nüsslein-Volhard and Wieschaus, 1980). The mutant phenotype observed in this study indicated the fundamental role of *hedgehog* in tissue patterning during embryonic development.

Later in the early nineties, the three vertebrate homologues of the *hedgehog* gene were identified; *Indian hedgehog* (*Ihh*), *Desert hedgehog* (*Dhh*) and *Sonic hedgehog* (*Shh*) (Echelard et al., 1993; Krauss et al., 1993; Riddle et al., 1993). Through amino acid sequence comparison, it was shown that the *Drosophila* hedgehog protein shared approximately 50% homology with its vertebrate counterparts (Echelard et al., 1993; Krauss et al., 1993; Riddle et al., 1993). Moreover, within the vertebrate hedgehog homologues the protein sequence was highly conserved with the N-terminus found to be almost identical in mouse, chick and zebrafish (Echelard et al., 1993). *Ihh* is an important regulator of skeletal development and is involved in chondrocyte

differentiation during embryogenesis (Vortkamp et al., 1996). *Dhh* regulates spermatogenesis in the testis (Bitgood et al., 1996). The role of *Shh* is discussed in more detail below.

1.3.1.2 *Shh* is a vertebrate morphogen

Of the three vertebrate homologues, *Shh* is the most widely expressed in the developing embryo. *Shh* expression in mouse can be detected in the central nervous system, limb buds and in the epithelial linings of the oral cavity to the hindgut (Figure 1.6) (Echelard et al., 1993). These tissues had been previously shown to have organisational properties, leading to the proposal that SHH acts as a morphogen which can pattern these tissues by inducing different cell fates in a concentration-dependent manner (Riddle et al., 1993).

Using homologous recombination to delete a 2.5kb portion of *Shh* in mouse embryonic stem cells that disrupts its signalling activity, Chiang et al were able to create a *Shh* knockout (*Shh^{KO}*) mouse model (Chiang et al., 1996). This mutation is embryonic lethal with defects observed from E8.5 including; a reduction in overall size of the embryo, lack of floorplate, midline, eye and distal limb structures as well as abnormal brain, gut, lung and craniofacial features (Chiang et al., 1996). This work indicates that not only tissues that express *Shh*, but also cells that respond to SHH signalling, are severely affected in the absence of SHH. These data supports a role of SHH as a morphogen during development whereby through its signalling pathway it can orchestrate broad tissue patterning across the embryo. SHH therefore has fundamental roles in establishing the body plan of developing vertebrate embryos.

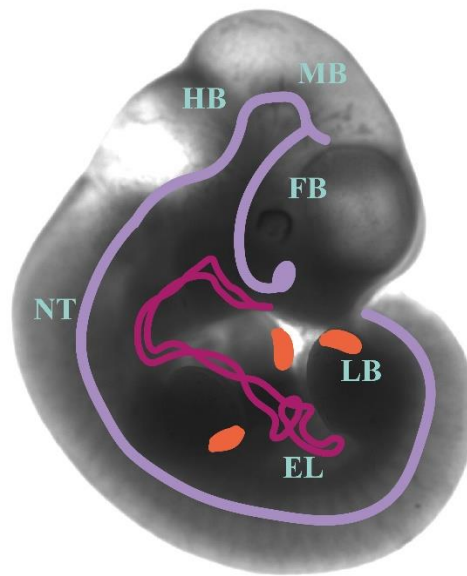


Figure 1.6. *Shh* is broadly expressed during embryonic development.

A schematic of murine *Shh* expression (coloured) in a developing E11.5 embryo. Expression of *Shh* can be detected throughout the central nervous system (purple) including the embryonic forebrain (FB), midbrain (MB), hindbrain (HB) and in the notochord and floorplate tissue of the developing neural tube (NT). *Shh* expression is also found in the zone of polarising activity of limb buds (LB - orange) and the epithelial linings (EL) of the oral cavity to the hindgut (pink).

1.3.2 SHH signalling pathway during development

SHH protein is a secreted ligand that acts within the highly conserved hedgehog signal transduction pathway. The full length 45 kDa precursor protein undergoes an autoproteolytic cleavage reaction to generate separate 25 kDa C-terminal (SHH-C) and 19 kDa N-terminal (SHH-N) domains (Figure 1.7A) (Lee et al., 1994). The SHH-C domain does not have any signalling functions but is required for cleavage of the precursor protein and the covalent attachment of a cholesterol molecule to the SHH-N (Figure 1.7A) (Porter et al., 1995, 1996b, 1996a). The covalent modification of the SHH-N protein has been proposed to be necessary for restricting the spatial activity of SHH-N (Porter et al., 1996b). The functional SHH-N domain is further processed, including palmitoylation of its N-terminus, before it can be secreted from cells and carry out both its short- and long-range signalling function in developing tissues (Hardy and Resh, 2012; Pepinsky et al., 1998; Porter et al., 1995).

Modified SHH-N is secreted from cells via Dispatched 1 (DISP1) where it can bind to a transmembrane receptor protein Patched 1 (PTCH1) (Figure 1.7B) (Burke et al., 1999; Marigo et al., 1996). In the absence of SHH binding, PTCH1 anchors another transmembrane protein called Smoothened (SMO) (Chen and Struhl, 1998; Murone et al., 1999). Upon SHH binding to PTCH1, SMO is released, activated and stabilised and is able to initiate a downstream signalling cascade that ultimately ends with the activation of glioma-associated (GLI) transcription factors (Figure 1.7B) (Chen and Struhl, 1998; Murone et al., 1999). The GLI transcription factors translocate to the nucleus of the SHH-responding cell and can exert gene expression changes (Figure 1.7B). The GLI family of transcription factors includes GLI1, which activates gene expression, and GLI2 and GLI3, both of which can act as activators or repressors of gene expression depending on their post-translation modifications (Hui et al., 1994; Ruiz i Altaba, 1997). Within mammals, much of the SHH signalling pathway occurs within a mammal-specific organelle termed the primary cilium (Murdoch and Copp, 2010). When SMO is released from PTCH1 it relocates to the primary cilium and initiates the signalling cascade to activate the GLI transcription factors within this structure (Figure 1.7B) (Murdoch and Copp, 2010).

SHH therefore mediates its signalling activity through regulation of its downstream effectors – the GLI family of transcription factors. Cells can respond to different concentrations of the morphogenic activity of secreted SHH and through both short- and long-range signalling activities induces different cell fates in neural and limb bud tissue (Briscoe et al., 2001; Riddle et al., 1993). SHH also displays a mitogenic function in limb buds where SHH-secreting cells are required for expansion of digital progenitors (Zhu and Mackem, 2011) and during late brain development where these promote cell proliferation needed for brain expansion (Britto et al., 2002). Additionally, SHH can act as a survival factor in myogenic precursor cells and is needed for limb muscle development (Krüger et al., 2001). The multi-faceted roles of SHH signalling are vital for embryonic development and disruptions to these pathways can severely impair normal developmental processes.

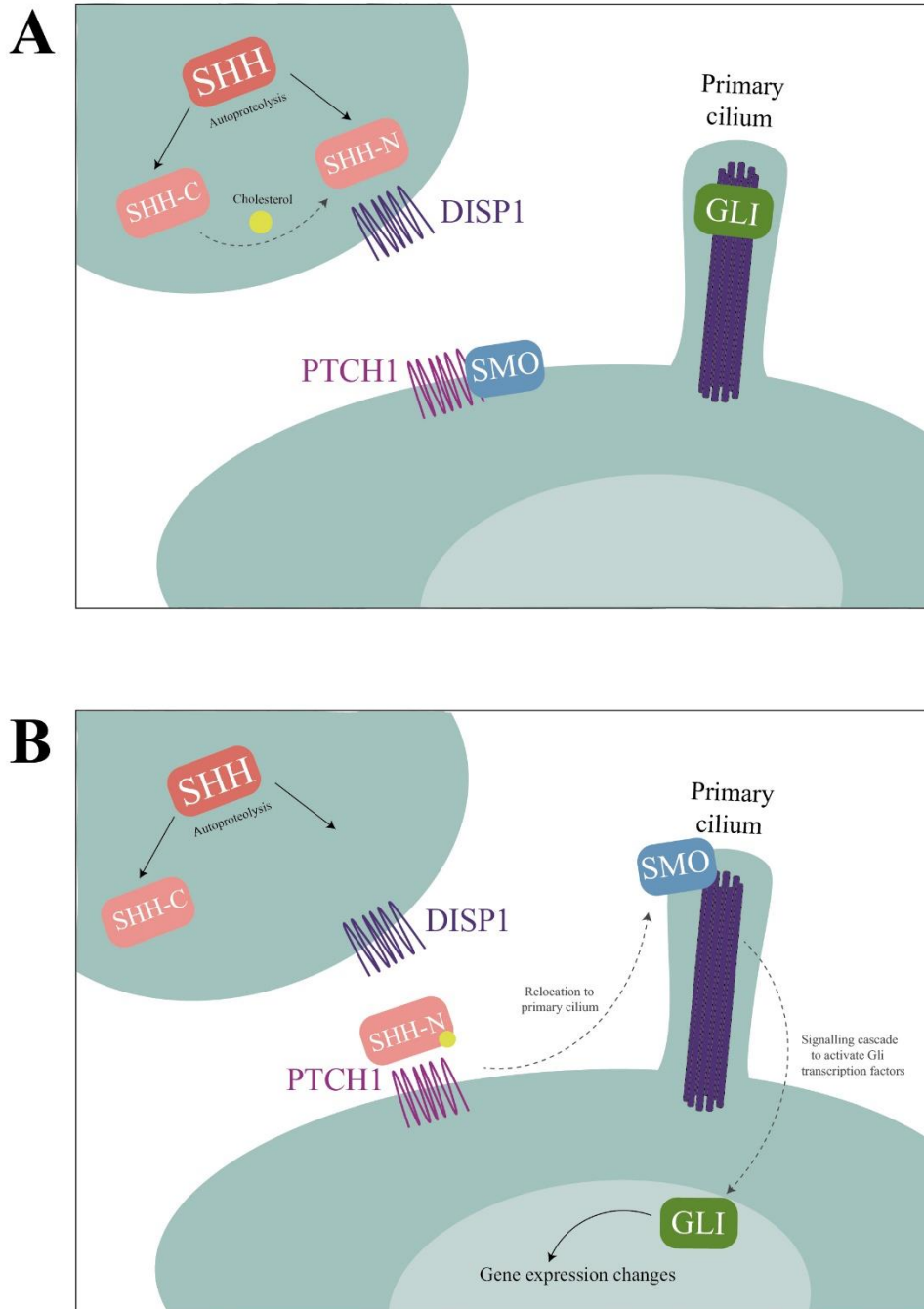


Figure 1.7. The SHH signalling pathway.

(A) The SHH precursor protein undergoes an autoproteolysis reaction to generate the SHH-C and SHH-N domains followed by the covalent attachment of cholesterol to SHH-N. (B) Modified SHH-N is secreted from cells via DISP1, binds to PTCH1 which causes the relocation of SMO to the primary cilium where activated SMO can initiate a signalling cascade to activate the GLI transcription factors that will ultimately lead to gene expression changes within the nucleus of the SHH-responding cell.

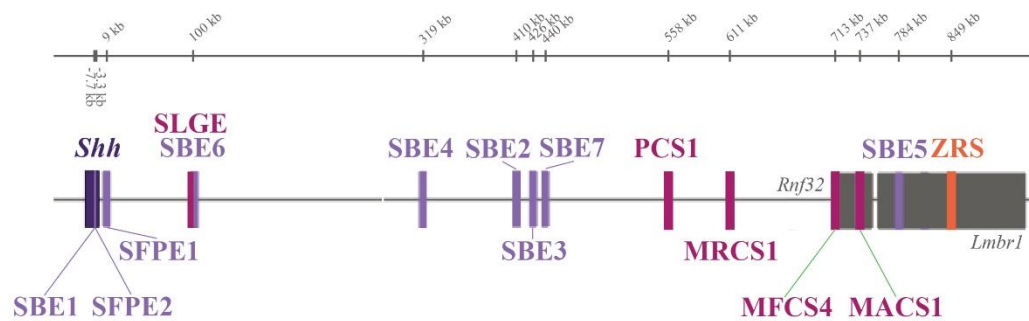
1.3.3 *Shh* has a complex regulatory landscape

The complex expression pattern observed for *Shh* during embryonic development is precisely controlled by an even more complex and intricate regulatory landscape. The *Shh* regulatory landscape is approximately 1 Mb in genomic distance with an 800 kb gene desert region found immediately upstream of *Shh* (Figure 1.8). So far, 15 tissue-specific enhancers of *Shh* have been described and the majority of these operate over large genomic distances (Figure 1.8A). The *Shh* enhancers can be found at various positions across the regulatory landscape including inside the *Shh* gene itself, within introns of other genes and throughout the gene desert region (Figure 1.8A). Some of these enhancers behave additively whereas others appear to have some, at least partial, redundant activity.

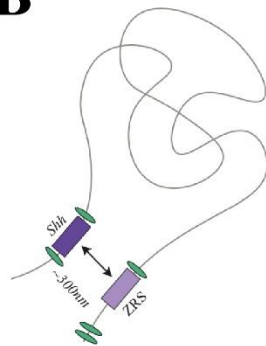
1.3.3.1 Limb bud expression

One of the most studied *Shh* enhancers is the limb-specific ZPA regulatory sequence or ‘ZRS’ enhancer (alternately termed MFCS1) (Lettice et al., 2003; Sagai et al., 2005). This is the most distally located enhancer of *Shh*, found 849 kb upstream in intron 5 of another ubiquitously expressed gene, *Lmbr1* (Figure 1.8A) (Lettice et al., 2003). This 780 bp element is highly conserved and has been demonstrated using a genetic *cis-trans* test and transgenic assays to be both necessary and sufficient for solely defining *Shh* expression in the zone of polarising activity (ZPA) in the distal posterior mesenchymal cells of the developing limb bud (Lettice et al., 2002, 2003; Sagai et al., 2004, 2005). The activity of the ZRS defines precise *Shh* levels in the limb bud through the binding of a balance of both activating and repressing transcription factors (Lettice et al., 2012). Point mutations that modify the binding of these factors cause ectopic *Shh* expression in the limb bud and cause congenital limb abnormalities (discussed further in 1.3.6). Interestingly, the ZRS is composed of two domains: a 5’ domain that directs limb bud-specific spatiotemporal activity of the enhancer and a 3’ domain that is required for long-range activity (Lettice et al., 2014).

A



B



C

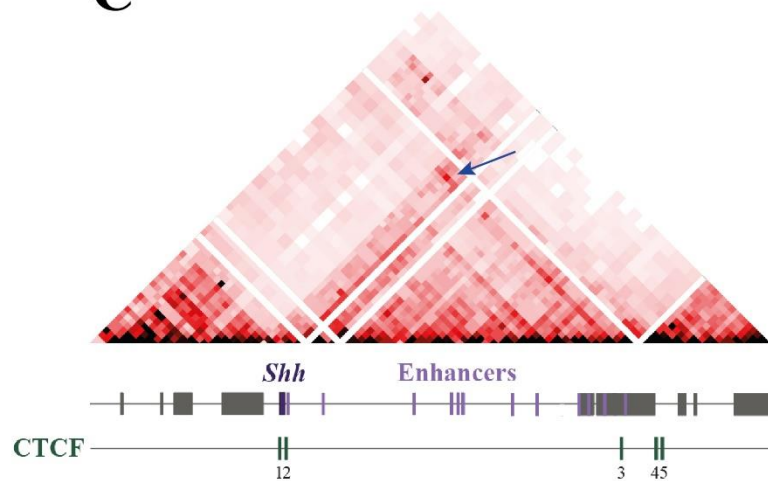


Figure 1.8 The *Shh* regulatory landscape.

(A) Schematic of the *Shh* regulatory landscape showing the position of genes (grey) and the *Shh* neural (lilac), limb (orange) and epithelial (pink) enhancers. The distance of each enhancer relative to the *Shh* promoter are displayed above. (B) Cartoon of the invariant chromatin loop that is believed to form between the *Shh* gene (purple) and the ZRS limb enhancer (lilac), mediated by CTCF (green). (C) The *Shh* TAD showing the peak interaction (blue arrow) that forms between *Shh* and ZRS. Below are tracks for genes (grey), enhancers (lilac) and the CTCF binding sites (green).

1.3.3.2 Central nervous system expression

Shh expression within the central nervous system has an indispensable function for ensuring proper brain, craniofacial and neural tube development during embryogenesis. Using a transgenic reporter assay to screen 35 kb of DNA surrounding the *Shh* gene, the functionally redundant *Shh* floorplate enhancers 1 and 2 (SFPE1 and SPFE2, respectively) were identified (Epstein et al., 1999). SFPE1 is 9 kb upstream of the *Shh* promoter whereas SFPE2 lies within intron 2 of *Shh*. These both drive reporter expression in the floorplate and hindbrain of the developing embryo (Figure 1.9) (Epstein et al., 1999) but their activation may be driven independently by different transcription factors (Jeong et al., 2006). More recently, *Shh* brain enhancer 6 (SBE6; 100 kb upstream) has been demonstrated to also drive reporter expression in the floorplate and hindbrain of developing embryos (Benabdallah et al., 2016).

This transgenic approach led to the identification of SBE1 within the second intron of *Shh* (Epstein et al., 1999). This enhancer directs *Shh* expression in the ventral midbrain and is required to maintain expression at later stages of embryogenesis (Figure 1.9) (Epstein et al., 1999; Jeong et al., 2011). A recent study which screened the *Shh* regulatory landscape for elements with a similar motif signature to SBE1 led to the discovery of SBE5 (784 kb upstream) (Yao et al., 2016). SBE5 is located in intron 15 of *Lmbr1* and is involved in *Shh* expression in the zona limitans intrathalamica (ZLI) of the midbrain but also displays partial redundancy with SBE1 in the ventral midbrain, ZLI and ventroposterior diencephalon (Figure 1.9) (Yao et al., 2016).

Employing, again, an enhancer trap assay to screen across 1 Mb of DNA encompassing *Shh* for regulatory potential combined with comparative sequence analysis, Jeong et al identified three forebrain enhancers in the gene desert region; SBE2 (410 kb upstream), SBE3 (426 kb upstream) and SBE4 (319 kb upstream). The activity of these enhancers is complex with some overlapping domains (Figure 1.9). SBE2 and SBE4 display activity in the ventral diencephalon but SBE2 is shown to be the only driver of LacZ expression in the rostral diencephalon (Jeong et al., 2006). SBE3 and SBE4 are both active in the ventral telencephalic region, with SBE3 activity restricted to subventricular zone of the medial ganglionic eminence (MGE), whereas SBE4 drives

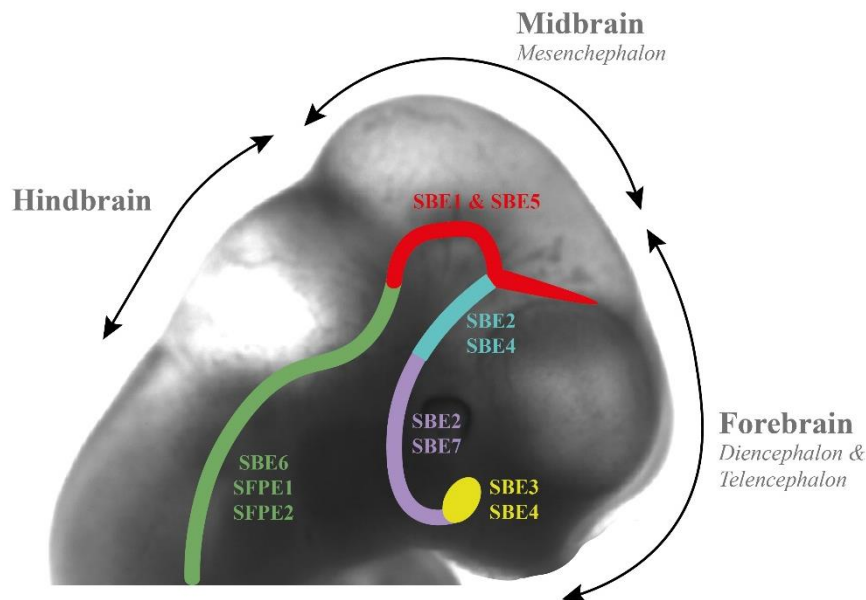


Figure 1.9. *Shh* regulation in neural tissue.

Schematic of neural *Shh* expression pattern (coloured) in a developing E11.5 embryo. The forebrain, midbrain and hindbrain regions are indicated. Domains corresponding to the activity of different enhancers are indicated by colour.

reporter expression in both the sub- and ventricular zone of the MGE (Jeong et al., 2006). A point mutation in SBE2 has been identified in a patient with holoprosencephaly (HPE), a brain malformation that results from the incomplete separation of the forebrain into the right and left hemispheres (discussed further in 1.3.6). This mutation was shown to compromise the ability of SBE2 to drive lacZ expression in the rostral diencephalon due to the abrogation of a Six3 binding site within the SBE2 enhancer (Jeong et al., 2008a). A subsequent study also found that this SBE2 mutation failed to drive reporter gene expression in the rostral diencephalon of zebrafish embryos (Bhatia et al., 2020).

By analysing ChIP-seq data of active enhancer marks from E14.5 brain tissue, Sagai et al identified SBE7 (~440 kb upstream) as a *Shh* enhancer of forebrain tissue and the precordal plate (PrCP) – a *Shh*-expressing mesendodermal tissue that is necessary for forebrain and craniofacial development. SBE7 shows redundant activities with the forebrain enhancers SBE2-4 and knockout mouse models of this enhancer displayed

midline defects reminiscent of HPE patients (Figure 1.9) (Sagai et al., 2019). Indeed, a recent study also demonstrated that in an SBE2 knockout mouse model, SBE7 (here termed forebrain secondary enhancer 1; FSE1) is responsible for the residual low levels of *Shh* observed in the forebrain of these mutants thereby preventing any developmental defects (Crane-Smith et al., 2019).

1.3.3.3 Epithelial tissue expression

Development of the epithelial linings of tissues such as the oral cavity and respiratory and digestive tubes is dependent on the expression of *Shh* (Litingtung et al., 1998). The three main enhancers that drive *Shh* expression in a co-linear manner within the epithelial linings of the oral cavity to the hindgut were discovered using a comparative sequence approach. Mammal reptile conserved sequence 1 (MRCS1; 611 kb upstream) drives reporter expression in the oral ectoderm, whereas mammal fish conserved sequence 4 (MFCS4; 713 kb upstream) showed reporter activity in the pharyngeal endoderm (Sagai et al., 2009). Mammal amphibian conserved sequence 1 (MACS1) is located in intron 8 of another gene, *Rnf32*, 737 kb upstream of *Shh* (Sagai et al., 2009). This enhancer has activity in the respiratory tube, gut endoderm as well as the urogenital tract. Genetic knockout mouse models of MFCS4 show post-natal lethality due to their inability to breathe or swallow properly demonstrating the requirement of this enhancer for the induction of *Shh* expression that leads to the development of pharyngeal structures (Sagai et al., 2009). *Shh* lung and gut enhancer (SLGE; 100 kb upstream) was identified using a comparative genomics approach in combination with subsequent assessment of the chromatin state of this candidate region (Tsukiji et al., 2014). SLGE-driven expression of *Shh* is detected in the oesophagus and stomach of developing embryos and shows redundant activity with MACS1. However, these enhancers show some distinct temporal activities with SLGE potentially contributing to *Shh* expression in the epithelial linings of adult tissue (Tsukiji et al., 2014). More recently, placental conserved sequence 1 (PCS1; 558 kb upstream) was shown to regulate *Shh* expression redundantly with MRCS1 in the oral cavity and whiskers of the developing embryo (Seo et al., 2018).

1.3.4 Do the *Shh* enhancers have an enhancer element signature?

As discussed in section 1.2.1.1, enhancers are defined here as a DNA sequence that shows increased transcription of its target gene in *cis* within its endogenous context when bound by cell type-specific TFs and with the associated chromatin marks.

The ZRS limb enhancer has been demonstrated to be necessary and sufficient for solely driving expression of *Shh* within the developing limb bud (Lettice et al., 2002, 2003; Sagai et al., 2004, 2005). The enhancer is also enriched for H3K4me1 and H3K27ac chromatin marks, a signature associated with active enhancers, specifically in the distal mesenchyme cells from the posterior side of the limb bud where *Shh* is expressed from the ZPA (Peluso et al., 2017). Using dissected limb bud cells from embryos, it has also been shown that the ZRS is bound by a combination of activating and repressing members of the ETS family of TFs (Lettice et al., 2012). Together these TFs define the pattern of *Shh* expression in the developing limb and mutations within these TF binding sites are associated with limb defects due to *Shh* misexpression (Lettice et al., 2012). When taken altogether, these work demonstrate that the ZRS has all the hallmarks of an enhancer according to the above definition.

In contrast, the majority of the neural enhancers (SFPE1/2, SBE1-7) have only been shown to drive reporter gene expression in a domain overlapping the *Shh* expression pattern rather than directly driving endogenous *Shh* expression (Benabdallah et al., 2016; Epstein et al., 1999; Jeong et al., 2006; Sagai et al., 2019; Yao et al., 2016). SBE6 drives reporter gene expression in neural progenitor cells, mouse and zebrafish brain tissue and is marked by H3K4me1 and H3K27ac modifications but knockout of this enhancer has no effect on *Shh* expression in neural progenitor cells (Benabdallah et al., 2016). The lack of impact on *Shh* expression following loss of these putative enhancers may be explained by the redundant activities of the *Shh* enhancers. In contrast, loss of enhancers SBE1, SBE5 or SBE7 reduces endogenous *Shh* expression in the midbrain, ZLI or forebrain, respectively, and these enhancers are marked by H3K27ac specifically in brain tissue (Paliou et al., 2019; Sagai et al., 2019; Yao et al., 2016). These neural enhancers (SBE1, SBE5 and SBE7) mostly carry the signature associated with an enhancer as defined above but more work is needed to identify the cell-type specific TFs that activate these enhancers.

The enhancers of *Shh* expression in the epithelial linings have been shown to be required for *Shh* expression in the pharynx (MFCS4) and teeth (PCS1) (Sagai et al., 2009; Seo et al., 2018). The majority of these enhancers have also only been shown to drive reporter gene expression in an overlapping domain with the *Shh* expression pattern rather than directly driving endogenous expression (Sagai et al., 2009; Seo et al., 2018; Tsukiji et al., 2014). These enhancers are also marked by H3K27ac in intestinal, lung and stomach tissue from E16.5 embryos (UCSC Genome Browser, Moore et al., 2020). More work is required to define the tissue-specific activities, chromatin features and associated TFs of these putative enhancers. One difficulty with studying the impact of a loss of a single enhancer on *Shh* transcription is that these enhancers often have overlapping domains of activity and loss of one enhancer can be compensated by another. One approach to circumvent issues when studying redundant enhancers is to knockout the enhancer in a sensitized background e.g. *Shh*^{+/-}. This has been employed previously to identify enhancers with overlapping activities and to detect any phenotypic abnormalities due to reduced expression (Osterwalder et al., 2018).

1.3.5 *Shh* functions in a three-dimensional genome

As discussed in 1.3.3, the majority of the *Shh* enhancers are found distal to the promoter within a mega-base regulatory landscape (Figure 1.8A). This landscape is a highly conserved region of synteny with the *Shh* enhancers found in the same order and orientation across different species (Goode et al., 2005; Woolfe et al., 2005). The limits of this regulatory landscape have been defined using transposon-based regulatory sensors (Anderson et al., 2014; Symmons et al., 2014, 2016). Unsurprisingly, the regulatory landscape of *Shh* overlaps with a TAD which contains the *Shh* gene and all of its known enhancers (Figure 1.8C). One of the TAD boundaries lies in an intragenic region that is 3' of *Shh*, whereas the other is found near the 5' end of *Lmbr1*. These reporter assays demonstrated that the boundaries of the *Shh* TAD mostly coincide with the limits of activity of *Shh* enhancers and appear to be invariant across cell types (Anderson et al., 2014; Symmons et al., 2016).

Using DNA FISH Amano et al. (2009), demonstrated that *Shh* looped out of its chromosome territory specifically in the *Shh*-expressing ZPA of developing limb buds

and this looping was dependent on the ZRS. *Shh* regulatory domain exists in a constrained compact chromatin state independent of *Shh* expression but increased *Shh* and ZRS colocalisation in the ZPA of limb cells, which may be consistent with a specific gene-enhancer looping interaction (Amano et al., 2009; Williamson et al., 2016). These studies show a direct link between higher order chromatin structure and gene regulation. The observation of *Shh* and ZRS being consistently found close together across all cell types and tissues examined so far implies that there may be an underlying invariant TAD structure that ensures distal enhancers communicate with the *Shh* promoter over large genomic distances (Figure 1.8B). Evidence from 3C studies show that the distal epithelial enhancers can also interact directly with the *Shh* promoter in a tissue-specific manner (Seo et al., 2018; Tsukiji et al., 2014). It remains unclear whether the *Shh* TAD is an invariant structure across the entire developing embryo as only a small number of tissues have been studied in any detail by DNA FISH or C-based approaches. Also, each enhancer of *Shh* has not been studied in its endogenous *in vivo* context and therefore, dynamic contacts between the *Shh* promoter and the relevant tissue-specific enhancer cannot be ruled out. A more recent study using DNA FISH and 5C found that activation of the neural enhancer SBE6 was accompanied by decreased enhancer-promoter spatial proximity (Benabdallah et al., 2019). This observation may be more consistent with a phase condensate model of gene regulation rather than the suspected chromatin looping between *Shh* and the ZRS. Whether the *Shh* locus uses different modes of enhancer-promoter communication or whether a unifying mechanism of distal regulation can be applied is still under investigation.

The *Shh* regulatory landscape contains at least five CTCF binding sites that are bound across multiple cell and tissue types clustered at the TAD boundaries in mouse (Figure 1.8C). This includes two strongly interacting convergent CTCF binding sites, which may act as loop anchoring points that form the *Shh* TAD by a loop extrusion mechanism (Fudenberg et al., 2016; Zuin et al., 2014). We speculated that these CTCF sites mediate the consistently close spatial proximity of *Shh* and ZRS (Figure 1.8B). Inversions that encompass the *Shh* TAD boundaries prevent the ZRS from regulating the *Shh* promoter, presumably by changing the CTCF site orientation, and blocks *Shh*-

ZRS interactions in limb bud tissue (Symmons et al, 2016). However, it has been difficult to study the *in vivo* function of CTCF during mammalian development due to the early embryonic lethality of CTCF null mice. Previous work, utilizing a conditional knockout of CTCF in the developing mouse limb, reported a loss of limb structure and a downregulation in *Shh* expression, but little effects on global gene expression patterns (Soshnikova et al., 2010). However, the *in vivo* role of CTCF of forming the TAD structure and potentially mediating gene regulatory interactions at the *Shh* locus remains unclear.

1.3.6 *SHH* misregulation can cause human developmental disorders

As discussed in 1.3.1, SHH is an essential protein for mammalian development and loss of the *Shh* gene is embryonic lethal. In humans, disruptions to the SHH signalling pathway and/or its tightly regulated expression pattern at key developmental stages can severely impair normal embryonic development and have pathogenic consequences. There are a number of human developmental disorders that have been identified caused by loss or mis-expression of *SHH* due to chromosomal rearrangements or mutations that disrupt *SHH* regulation via its distal enhancers.

One of the most studied human congenital malformations related to *SHH* is a dominant, highly penetrant condition called preaxial polydactyly (PPD). The most frequent cause of PPD (1:2000 births) was mapped to a region of human chromosome 7 from which patients present with a range of limb phenotypes from triphalangial thumb (thumb converted to finger) to PPD (extra digits) (for review see e.g. Hill, 2007). The orthologous region on mouse chromosome 5 had also been implicated in PPD and other limb defects (Lettice et al., 2002, 2003; Sagai et al., 2004). Both mouse and human genetic studies identified intron 5 of *LMBR1* as being the key region disrupted in PPD patients. As the PPD phenotype was physically linked to the *SHH* gene, approximately 1 Mb away, this led to the proposal that disruption of a long-range regulator of *SHH* was the underlying genetic mechanism of PPD (discussed in 1.3.3.1). Subsequent studies demonstrated that gain of function mutations in the ZRS act in *cis* and cause anterior ectopic expression of *SHH* in the developing limb bud (reviewed by e.g. Anderson et al., 2012). Point mutations in the ZRS disrupt transcription factor

binding sites through both gain and loss of binding motifs and as a consequence the *Shh* expression domain in limb buds (Lettice et al., 2012). Another mutation found within the *SHH* regulatory landscape is found in patients with a rare recessive condition called acheiropodia (Ianakiev et al., 2001). The patients display bilateral limb truncations and the mutation has been mapped to a region about 30 kb upstream of the ZRS (Ianakiev et al., 2001). The phenotype of these patients is similar to limb defects observed in *Shh* and ZRS knockout mouse models (Chiang et al., 1996; Lettice et al., 2014; Sagai et al., 2005). This has led to the proposal that the acheiropodia mutation disrupts the long-range regulation of *SHH* by ZRS; however, the disease mechanism of this disorder is still unclear. Collectively, these limb malformations are termed ZRS-associated syndromes (Wieczorek et al., 2010).

SHH has also been identified as one of the most common genetic causes for HPE, a disorder characterised by failure of the forebrain to divide resulting in midline defects with variable expressivity including cyclopia and cleft palate/lip (Dubourg et al., 2007; Hehr et al., 2004; Roessler et al., 1996). Heterozygous *SHH* mutations segregate with HPE patients, whereas in mice, loss of both copies of *Shh* are necessary to recapitulate a HPE-like phenotype (Chiang et al., 1996; Roessler et al., 1996). The haploinsufficiency of *SHH* observed in HPE patients suggests that humans may be more sensitive to SHH levels in the developing forebrain and prechordal plate than mice and highlights a species-specific difference in *SHH* regulation. HPE patients that carry chromosomal rearrangements that disrupt a region upstream of *SHH* were also identified (Belloni et al., 1996). These rearrangements likely disrupt *SHH* regulation by displacing the forebrain enhancers SBE2/3/4/7 from the *SHH* promoter (Jeong et al., 2006; Sagai et al., 2019). An HPE patient has been identified with a point mutation in SBE2 that disrupts a Six3 binding sites and has reduced reporter gene activity in the rostral diencephalon (Jeong et al., 2008a). This study indicates that loss of SBE2 activity through reduced transcription factor binding in the developing forebrain can cause HPE by reducing *SHH* expression during embryogenesis (Jeong et al., 2008a).

In summary, mutations to the *SHH* enhancers that impact precise spatio-temporal expression patterns by changing TF binding sites or rearrangements that disrupt the

regulatory landscape can cause severe developmental disorders. It is important that we understand what trans-acting factors are required for enhancer function and how these long-range interactions are formed within the genome to better understand the underlying disease mechanisms.

1.4 Aims of PhD thesis

The overall aim of my thesis is to further understand how distal enhancers operate over large distances to regulate their target gene during mammalian development. To do this, I used *Shh* as a model locus to study the spatiotemporal dynamics of distal enhancers and the mechanisms of long-range gene regulation in relation to three-dimensional chromatin structure. The objectives of my thesis include:

Objective 1: Understand the spatiotemporal dynamics of *Shh* neural enhancers during embryonic brain development.

Objective 2: Study the developmental role of CTCF binding sites in *Shh* TAD structure and long-range gene regulation.

Objective 3: Use a synthetic transcription factor system to test the importance of *Shh* TAD structure for long-range enhancer-driven gene activation.

Objective 4: Analyse the chromatin structure of the *SHH* TAD in acheiropodia patient-derived cells carrying a human-specific CTCF binding site deletion.

Chapter 2. Materials & Methods

2.1 General molecular biology reagents

Table 2.1 General molecular biology reagents used.

Reagent	Supplier	Cat. number
Agarose	BioGene	300-300
Bovine serum albumin (BSA)	Sigma-Aldrich	A3294
Dextran sulphate	Fisher Scientific	BP1585-100
Diethyl pyrocarbonate (DEPC)	Sigma-Aldrich	D5758
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich	D2650
Ethanol (EtOH)	Fisher Scientific	E/0650DF/17
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich	E9884
Formamide	Sigma-Aldrich	F7503
Gelatin	Sigma-Aldrich	G1890
Glycerol	Fisher Scientific	G/0600/08
IGEPAL CA-360	Sigma-Aldrich	F7503
Isopropanol	Fisher Scientific	326961000
Paraformaldehyde (PFA)	Sigma-Aldrich	16005
Phenol-chloroform	Sigma-Aldrich	S77617
Proteinase K	Sigma-Aldrich	P8044
Triton X-100	Sigma-Aldrich	T8787
Tween 20	Sigma-Aldrich	P1379
Xylene	Fisher Scientific	X/0250/17

2.2 Cell culture

2.2.1 Mouse embryonic stem cells

Mouse embryonic stem cells (mESCs) were grown on 0.1% gelatin (Sigma-Aldrich G1890) coated Corning flasks in GMEM BHK-21 (Gibco 21710-025) media supplemented with 15% FCS, 1000 units/ml Leukemia inhibitory factor (LIF; produced in-house), 2 mM L-glutamine, 1 mM sodium pyruvate (Sigma-Aldrich 58636), 5X non-essential amino acids (Sigma-Aldrich M7145) and 50mM 2- β -mercaptoethanol (Gibco 31350-010). Cells were passaged at 60-90% confluence by washing with PBS, treating with trypsin (0.05% v/v; Gibco 25300-054) for ~2 minutes (mins) at 37°C and tapping flasks to detach cells. Trypsin was inactivated by addition of ten volumes of complete media and the mixture was pipetted repeatedly to generate a single-cell suspension. Cells were centrifuged at 1200 rpm for 4 mins and counted using a haemocytometer. Cells were plated onto a fresh flask at a density of approximately 4×10^4 cells/cm². Fresh media was added on the days cells were not passaged. Cells were maintained at 37°C with 5% CO₂. For long-term storage of cell lines in liquid nitrogen, approximately $0.5-1 \times 10^6$ cells were frozen down in complete media supplemented with 10% DMSO and 10% FCS.

The mESCs used here include wild type E14 (both in-house lines and the parental line of the CTCF-AID cells provided by Elphege Nora), CTCF-AID (Nora et al., 2017) and Scc1-AID (Rhodes et al., 2020). Unless specified, in-house wild type E14 cells were used in mESC experiments.

2.2.2 Human lymphoblastoid and fibroblast cell lines

Human lymphoblastoid cells were grown on Corning flasks in IMDM media (Gibco 12440053) supplemented with 10% FCS, 2 mM L-glutamine and P/S. Cells were passaged and frozen as described in 2.2.1. Cells were maintained at 37°C with 5% O₂.

Human fibroblast cells were cultured by Patricia Heyn (Andrew Jackson Research Group). Glass slides were prepared for DNA FISH as described in 2.7.4 following cell fixation and permeabilisation by Patricia.

2.3 Mouse tissue

2.3.1 Embryonic tissue dissection

All mouse lines described throughout were housed either in the Biomedical Research Facility (semi-barrier mouse facility) or Transgenic Unit (full barrier mouse facility). At an appropriate stage, uteri were collected by facility staff. Embryos were dissected out of uteri in ice-cold PBS, washed in PBS and then fixed in 4% PFA overnight at 4°C. Embryos were stored at 4°C in either PBS or 70% EtOH.

2.3.2 Genotyping

The majority of mouse line genotyping was carried out by Lorraine Rose or Laura Lettice. For genotyping of embryos, either tail or limb bud tissue was removed and incubated in 500 µl lysis buffer (100 mM Tris pH 8.5, 5 mM EDTA, 200 mM NaCl and 0.2% SDS) and 10 µg/mL proteinase K overnight at 55°C. To extract DNA, samples were briefly vortexed, an equal volume of isopropanol was added and then placed at -80°C for ~1 hour until DNA pellet formed. Supernatant was removed, DNA pellet was washed in ~200 µL of 70% EtOH and then allowed to air dry. DNA pellet was resuspended in 500 µL of TE buffer and stored at -20°C.

For PCR of extracted DNA, 25 µL reactions were prepared using the dNTP set (Invitrogen 10297-117) and components of the Platimun Taq DNA polymerase set (Invitrogen 10966-026) according to the manufacturer's instructions. PCR was performed on a thermal cycler (Table 2.2) and reaction products were run on a 2% agarose gel for analysis of genotype.

Table 2.2 Genotyping thermal cycler programme.

Temperature (°C)	Time (seconds)	Cycles
94	300	
94	30	34
58	30	
72	90	
4	∞	

2.3.3 Tissue embedding and sectioning

To embed tissue in paraffin wax, previously fixed embryos were transferred through the following series; 70% EtOH, 90% EtOH, 100% EtOH, 100% EtOH, xylene, xylene (60°C), xylene: wax (50:50, 60°C) and 2 x wax (60°C). All incubations were for ~10-20 mins at room temperature unless stated otherwise. Embryos were then transferred to a mold, covered with wax, repositioned into the correct orientation for sectioning and left to cool and set at -4°C for minimum 1 hour. Embedded embryos were sectioned at 6-8 μ M on a microtome and collected on SuperFrost Plus Adhesion glass slides (Fisher Scientific 10149870). Tissue sections were stored at room temperature.

2.4 Assay for transposase-accessible chromatin

2.4.1 Mouse model information

ShhGFP/cre mice express a fusion product of GFP and Cre recombinase from the endogenous *Shh* locus (N.B. this creates a null allele of *Shh*). All *Shh*-expressing cells in these mice will also express GFP and Cre recombinase.

Membrane-targeted tdTomato/ membrane-targeted EGFP (mTmG) mice have a dual fluorescent Cre-reporter allele. In the absence of Cre recombinase, all cells will express membrane-targeted tdTomato (mT) and will appear as red fluorescence. In the presence of Cre recombinase, loxP sites found at either side of the mT cassette are targeted leading to the deletion of the mT cassette causing cells to switch to expression of membrane-targeted EGFP (mG) and will appear as green fluorescence.

Crossing ShhGFP/cre to mTmG strains generates compound heterozygote ShhGFP/cre;mTmG embryos meaning that in *Shh*-expressing cells, the expression of Cre recombinase causes a deletion of the mT cassette and subsequent expression of membrane-targeted EGFP (mG). Thereby, we can distinguish between *Shh*-expressing (mG) and non-expressing (mT) cell populations in these ShhGFP/cre;mTmG embryos by green and red fluorescence.

2.4.2 Preparation of single-cell suspensions

Embryos were collected and dissected out as described (2.3.1). GFP positive (GFP+) embryos were identified using a fluorescent microscope. E10.5, E11.5 and E14.5 GFP+ embryos were cut between the maxillary and mandibular processes across to the midbrain/hindbrain junction. Single cells were obtained by incubating dissected tissue in 100 μ L trypsin at 37°C for 15 minutes. Media was added to stop reaction followed by cell pelleting by centrifugation (500g for 5 mins at 4°C). Cell pellet was washed and resuspended in 500 μ L cold PBS.

2.4.3 Fluorescent-activated cell sorting

Samples were taken to the FACS facility for sorting of GFP and RFP cells from mouse tissue using a BD FACSAriaII SORP cell sorter. Data was acquired using BD FACSDive. GFP fluorescence was detected using a blue 488nm excitation laser and a 530/30nm bandpass emission filter. RFP fluorescence was detected using a yellow-green 561nm excitation laser and a 582/15nm bandpass emission filter. Gates on FSC-H/FSC-A and on SSC-A/SSC-H were used to eliminate clumps and debris. The sort gates for GFP+ and RFP+ subpopulations were applied on cells from non-expressing *Shh* tissue of the same sample.

ShhGFP/cre samples were sorted into GFP+ and GFP- populations and ShhGFP/cre;mTmG samples were sorted into GFP+ and RFP+ populations.

2.4.4 Transposition reaction

Sorted cells were pelleted (500g for 5 mins at 4°C), washed in cold PBS, then gently resuspended in 50 μ L cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl₂ and 0.1% IGEPAL CA-360) and pelleted at 500g for 10 mins at 4°C. Supernatant was discarded and cell pellets placed on ice. Cells were gently resuspended in 50 μ L transposition reaction mix as follows: 2X Tagment DNA buffer (Illumina 15027866), 20X Tagment DNA enzyme (Illumina 15027865) in ddH₂O. Samples were incubated at 37°C for 30 mins. The transposed samples were then purified using the MinElute PCR purification kit (Qiagen 28204), according to the manufacturer's instructions, with a final elution volume of 10 μ L. Purified DNA samples were stored at -20°C.

2.4.5 ATAC library amplification

Transposed DNA samples were amplified by PCR by setting up a 50 µL reaction as follows: 10 µL transposed DNA, 9.7 µL ddH₂O, 2.5 µL 25 µM customised Nextera PCR primer 1, 2.5 µL 25 µM customised Nextera PCR primer 2* (Table 2.3), 0.3 µL 100X SYBR Green I (Invitrogen S-7563) and 25 µL 2X NEBNext high-fidelity PCR master mix (New England Biosciences M0541). PCR amplification was performed on a thermal cycler (for programme details see Table 2.4).

In order to reduce GC/size bias and over amplification of libraries during PCR, the reaction was monitored by quantitative PCR (qPCR) to stop amplification prior to saturation. A qPCR 15 µL side-reaction was set up as follows: 5 µL 5 cycles PCR amplified DNA, 4.44 µL ddH₂O, 0.25 µL 25 µM customised Nextera PCR primer 1, 0.25 µL 25 µM customised Nextera PCR primer 2* (Table 2.3), 0.06 µL 100X SYBR Green I and 5 µL 2X NEBNext high-fidelity PCR master mix. qPCR was performed on a Lightcycler 480 (for programme details see

Table 2.5). The additional number of cycles required for each sample was determined by plotting a linear run versus the cycle number. The number of cycles that correspond to ¼ maximum fluorescent intensity were calculated for each sample. The remaining 45 µL 5 cycles PCR amplified DNA was run as before (Table 2.4, omitting 72°C initial step and modifying the number of cycles to the calculated amount). Amplified DNA samples were stored at -20°C.

Table 2.3. Nextera PCR primers used for library amplification.

Nextera PCR primer	Sequence
1	AATGATACGGCGACCAACGAGATCTACACTCGTC GGCAGCGTCAGATGTG
2	CAAGCAGAAGACGGCATAACGAGATTTCGCCTTAG TCTCGTGGGCTCGGAGATGT
3	CAAGCAGAAGACGGCATAACGAGATCTAGTACGG TCTCGTGGGCTCGGAGATGT
4	CAAGCAGAAGACGGCATAACGAGATTTCTGCCTG TCTCGTGGGCTCGGAGATGT
5	CAAGCAGAAGACGGCATAACGAGATGCTCAGGAG TCTCGTGGGCTCGGAGATGT
6	CAAGCAGAAGACGGCATAACGAGATCATGCCTAG TCTCGTGGGCTCGGAGATGT
7	CAAGCAGAAGACGGCATAACGAGATGTAGAGAGG TCTCGTGGGCTCGGAGATGT
8	CAAGCAGAAGACGGCATAACGAGATCCTCTCTGG TCTCGTGGGCTCGGAGATGT

*PCR primer 1 was used as forward primer for all samples whereas primers 2-8 were used as barcoded reverse primers in order to pool multiple samples per sequencing lane.

Table 2.4. Library amplification thermal cycler programme.

Temperature (°C)	Time (seconds)	Cycles
72	300	
98	30	
98	10	4
63	30	
72	60	
4	∞	

Table 2.5. qPCR of amplified libraries thermal cyclers programme.

Temperature (°C)	Time (seconds)	Cycles
98	30	
98	10	19
63	30	
72	60	
4	∞	

2.4.6 ATAC library purification

Amplified DNA samples were subjected to double size selection to remove DNA fragments <150 base pairs (bp) and >1000 bp that would hinder sequencing reactions. Samples were made up to 90 µL with ddH₂O and purified using the SPRIselect beads (Beckman Coulter B23317), according to the manufacturer's instructions. For removal of large DNA fragments, a ratio of 0.55 was used, and for removal of small DNA fragments, a ratio of 1.8 was used. Purified DNA samples were eluted in 20 µL TE buffer and stored at -20°C.

2.4.7 ATAC library quality control and sequencing

DNA samples were quantified using the Qubit dsDNA high sensitivity assay (ThermoFisher Q32854) with the Qubit 4 fluorometer. The quality of DNA samples were analysed using the high sensitivity DNA kit (Agilent 5067-4626) with the Bioanalyzer instrument by the IGMM Technical Services. Samples with the nucleosomal fragment distribution profile expected for ATAC-seq libraries (mono-, di- and trisomal fragments) were sent for sequencing. Pooled ATAC-seq libraries were subjected to 50 bp paired end sequencing using the HiSeq 4000 platform at BGI in Hong Kong.

2.5 ATAC-seq computational analysis

2.5.1 Read mapping

FastQC was used to obtain basic quality control metrics from sequencing data and assess quality of reads before and after preprocessing. Sequencing reads were trimmed to a minimum of 30 bases followed by adaptor sequences clipping using cutadapt. Reads were aligned to mouse genome assembly mm9 using bowtie2. Mitochondrial reads (using awk) and PCR (with Picard) duplicates were filtered out.

2.5.2 Quality control

Fragment length distribution of sequencing reads were obtained. Alignment statistics were collected using samtools idxstats. Principle component analysis (PCA) was performed to compare all samples.

2.5.3 Peak calling and analysis

Reads were shifted by +4 bp for the positive strand and -5 bp for the negative strand. Broad and narrow peaks were called using MACS2 callpeak. Irreproducible discovery rate (IDR) analysis was performed to assess the consistency between peaks called per replicate. Bedgraph output files from MACS2 PE narrow mode were converted to bigWig format using UCSC utilities and tools for visualisation of data on UCSC. Sample replicates were merged then GFP+ and GFP- cell populations were intersected using BEDtools. Peaks unique to either GFP+ or GFP- cells could be extracted from each developmental stage. Gene ontology (GO) analysis was performed using Cistrome-GO (<http://go.cistrome.org/>) with the following parameters: solo mode, 10,000 top peaks, FDR of 0.05% and a minimum of 200 and maximum of 2000 genes in GO and KEGG gene sets. Motif analysis was performed using HOMER (<http://homer.ucsd.edu/homer/motif/>).

2.6 Immunofluorescence

2.6.1 Immunofluorescence solutions

Details of the solutions used for immunofluorescence of mouse embryonic tissue are found in Table 2.6.

Table 2.6. Immunofluorescence solutions.

Buffer	Composition
TEG buffer	2.422g TRIS(hydroxymethyl) methylamine and 0.38g EGTA made up 2 L with ddH ₂ O (pH 9)
Ammonium chloride (NH ₄ Cl) buffer	2.6745g NH ₄ Cl made up to 1 L with PBS
Wash 1 buffer	2g gelatin (0.2%), 10g BSA (1%) and 0.5g saponin (0.05%) made up to 1 L with PBS
Wash 2 buffer	4g gelatin (0.2%), 2g BSA (0.1%) and 1g saponin (0.05%) made up to 2 L with PBS
Antibody solution	0.05g BSA and 150 Triton X-100 made up to 50 mL with PBS

2.6.2 Immunofluorescence of mouse embryonic tissue

Alternate slides carrying sections through mouse embryos were selected to identify regions of Shh expression by immunofluorescence. Paraffin embedded tissue sections were incubated at 50°C for 2-24 hours. Slides were dewaxed and rehydrated by transferring through the following series; 3 x xylene, 3 x 100% EtOH, 90% EtOH, 70% EtOH and washed 3-4 times in ddH₂O. All incubations were for ~10-20 mins at room temperature.

Slides were subjected to antigen retrieval using TEG buffer for 4 mins in a pressure cooker. Following antigen retrieval, slides were cooled in TEG buffer for a minimum of 30 mins and then washed in NH₄Cl buffer for 30 mins. Slides were blocked in three 10 minute washes in Wash 1 buffer. Slides were incubated with primary antibody raised against the C-terminus of Shh (Abcam ab86462), diluted 1:150 in antibody

solution, overnight at 4°C in a humidified chamber. Slides were allowed to reach room temperature for a minimum of 30 mins and then washed three times in Wash 2 buffer (10 mins/wash). Slides were incubated with secondary antibody (Alexa-Fluor 488, Invitrogen A-212208) diluted 1:400 in antibody solution for 1 hour at room temperature and then washed three times in Wash 2 buffer. Slides were washed once in PBS, stained with 0.5 µg/mL DAPI (Sigma D9542) and mounted using Vectashield (Vector laboratories H-1000). Slides were stored at 4°C.

2.7 DNA fluorescent in situ hybridisation

2.7.1 Fosmid probe preparation

Fosmid probes (Whitehead, Sanger CHECK***) were streaked onto chloramphenicol (12.5 ng/µL) agar plates and incubated overnight at 37°C. Individual clones were picked and incubated overnight shaking at 225 rpm in 5 mL L-broth. Glycerol stocks of fosmid probes were made by added 170 µL of glycerol to 830 µL fosmid culture and stored at -80°C. For details of fosmids used see Table 2.7.

To isolate fosmid probe DNA, a pipette tip amount of fosmid glycerol stock was incubated overnight at 37°C shaking at 225 rpm in 5 mL L-broth containing chloramphenicol (12.5 ng/µL). Fosmid cultures were pelleted by centrifugation (10 mins at 4000 rpm) and supernatant was discarded. Cell pellet was resuspended in 200 µL GTE buffer (50 mM glucose, 25 mM Tris pH 8 and 10 mM EDTA) containing lysozyme and 1 µL RNase A (20 mg/mL; Invitrogen 12091021), mixed vigorously to ensure pellet was well resuspended and then incubated at room temperature for 5 mins. 400 µL lysis buffer (0.2 M NaOH, 1% SDS) was added, mixed by inversion and then incubated on ice for 5 mins. To stop the reaction 300 µL acetate buffer (5M potassium acetate, glacial acetic acid) was then added, gently vortexed and incubated on ice for 5 mins. Samples were centrifuged at maximum speed for 5 mins at 4°C. Fosmid DNA was purified by phenol: chloroform extraction, pelleted by centrifugation (maximum speed for 15 mins at 4°C), washed in 70% EtOH, dried and resuspended in 25 µL TE buffer. Fosmid miniprep DNA was quantified using the Qubit dsDNA broad range assay (ThermoFisher Q32850) with the Qubit 4 fluorometer and then stored at -20°C.

Table 2.7. Details of fosmid probes.

Species	Region	Sanger ID	Whitehead ID	Coordinates
Mouse	Shh	G135P64333A4	WI1-574O18	chr5:28754458-28795879
Mouse	SBE2	G135P603171G8	WI1-1275C09	chr5: 29195832-29239355
Mouse	ZRS	G135P600929F6	WI1-1047E14	chr5: 29611727-29653695
Human	SHH	G248P82324G11	WI2-0444K07	chr5: 155585515-155620891
Human	LSC1	G248P83074E9	WI2-1039I18	chr7: 156248055-156288749
Human	LSC2	G248P80007C10		chr7: 156537478-156580173
Human	LSC3-5	G248P8090C7	WI2-0544E14	chr7: 156591166-156630143
Human	Mnx1	G248P84315E3	WI2-1461J6	chr7: 156785169-156823804

2.7.2 Fosmid probe labelling

Fosmid DNA was labelled by a nick translation reaction in a 20 μ L reaction mix as follows: 6 μ L of fosmid miniprep DNA (0.5 – 1 μ g), 2 μ L nick translation salts (0.5 M Tris pH 7.5, 0.1 M MgSO₄, 1 mM DTT and 0.5 mg/ml BSA), 1 μ L DNase I (1:5 dilution in ice-cold H₂O, Roche 04716728001), 1 μ L DNA polymerase I (Invitrogen 18010-017), 2.5 μ L 0.5 mM dATP, 2.5 μ L 0.5 mM dCTP, 2.5 μ L 0.5 mM dGTP and 2.5 μ L of labelled dUTP (Table 2.8). Reaction mixes were mixed, centrifuged briefly and then incubated at 16°C for 1.5 hours. Nick translation reaction was stopped by the addition of 3 μ L 0.5 M EDTA and 2 μ L 20% SDS. Sample volume was made up to 90 μ L with TE buffer and then purified using Quick Spin Columns (Roche 11273973001)

as described by the manufacturer's instructions. Labelled fosmid probes were stored at -20°C.

Table 2.8. dUTPs used to label fosmid probes.

Labelled dUTP	Type of label	Product details
Green496-dUTP	Direct green	ENZO lifesciences ENZ-42831L
ChromaTide Alexa Fluor 594-5-dUTP	Direct red	Thermofisher scientific C11400
Aminoallyl-dUTP	Direct far red	Jena bioscience NU-803-647N-S
Bio-16-dUTP	Biotin	Roche 11093070910
Digoxigenin-11-dUTP*	Digoxigenin	Roche 3359247910

*1.5 µL Digoxigenin-11-dUTP and 1 µL 0.5mM dTTP were used per reaction labelled with digoxigenin.

2.7.3 3D FISH of tissue sections

Paraffin embedded tissue sections were incubated at 60°C for 20 mins. Slides were dewaxed and rehydrated by transferring through the following series; 3 x xylene, 3 x 100% EtOH, 2 x 95% EtOH, 2 x 70% EtOH (5 mins) and final wash in ddH₂O. All incubations were for ~10 mins at room temperature unless stated otherwise. Slides were microwaved in 0.1 M citrate buffer (pH 6) for 60 mins, allowed to cool in buffer for a minimum of 20 mins and then washed in ddH₂O. Slides were rinsed in 2X SSC and then incubated in 2X SSC at 75°C for 5 mins. Slides were denatured in 70% formamide/ 2X SSC at 75°C for 3 mins followed by 3 mins in ice-cold 70% EtOH and then dehydrated by transferring through 90% EtOH then 100% EtOH (2 mins each). Slides were air dried.

Approximately 80-100 ng of labelled fosmid probe was applied per slide by preparing with 6 µg mouse Cot-1 DNA (Invitrogen 18440016) per fosmid, 5 µg salmon sperm DNA (Invitrogen 15632011) per slide in two volumes of EtOH. Fosmid probe reaction mix was pelleted, dried and resuspended in 12 µL hybridisation mix (50% formamide, 10% dextran sulphate, 2X SSC, 1% Tween 20 in ddH₂O). Probes were left to dissolve in hybridisation mix for a minimum of 1 hour, vortexed and centrifuged briefly,

denatured at 75°C for 5 mins, reannealed at 37°C for 15 mins and hybridised to the denatured slides at 37°C for 24-72 hours.

Slides were washed four times for 3 mins in 2X SSC at 45°C, four times for 3 mins in 0.1X SSC at 60°C and then transferred to 4X SSC containing 0.1% Tween 20. For biotin and digoxigenin labelled probes, slides were incubated with blocking buffer (4X SSC with 5% Marvel) at room temperature for 5 mins. Antibodies used to detect biotin and digoxigenin labelled probes were diluted in blocking buffer and centrifuged at maximum speed for 15 mins at 4°C (for antibody details see Table 2.9). Slides were incubated with antibodies for 1 hour at 37°C in a humidified chamber with three washes for 2 mins in 4X SSC containing 0.1% Tween 20 between different antibody incubations. Slides were stained with 0.5 µg/mL DAPI, mounted using Vectashield and stored at 4°C prior to imaging.

Table 2.9. Antibodies used to detect indirectly labelled fosmid probes.

Antibody	Concentration	Product details
Biotinylated anti-avidin	1:100	Vector labs
Streptavidin Cy5	1:10	Amersham
Texas Red anti-sheep	1:100	**CHECK
Rhodamine anti-DIG FAB fragments	1:20	Roche

2.7.4 3D FISH of cell lines

0.5-1 x 10⁶ cells were seeded onto autoclaved SuperFrost Plus Adhesion glass slides (mESCs and human fibroblasts; Fisher Scientific 10149870) or Poly-prep slides (Human lymphoblastoid cells; Sigma P0425) overnight. Cells were washed in PBS, fixed in 4% PFA for 10 mins, washed again in PBS and then permeabilised using 0.5% Triton X for 10 mins. Cells were left overnight in PBS at 4°C, dried and then stored at -80°C prior to DNA FISH.

Slides were washed briefly in PBS, 2X SSC and then incubated in 2X SSC containing 100 µg/mL RNase (Invitrogen 12091021) for 1 hour at 37°C. Slides were rinsed in 2X SSC and then dehydrated by transferring through 70% EtOH, 90% EtOH, 100% EtOH

(2 mins each) and then air dried. Slides were incubated at 70°C for 5 mins and then denatured in 70% formamide/ 2X SSC at 80°C for 15-60 mins (depending on cell type) followed by 3 mins in ice-cold 70% EtOH and then dehydrated by transferring through 90% EtOH then 100% EtOH (2 mins each). Human cells were denatured for 30 mins and mouse cells for 40 mins. Slides were air dried.

Fosmid probe preparation, hybridisations, washes, antibody incubations and mounting of slides was carried out as described in 2.7.3.

2.8 RNA fluorescent in situ hybridisation

2.8.1 RNA FISH of tissue sections

Custom Stellaris® RNA FISH Probes were designed against *Shh* nascent mRNA (pool of 48 unique 22-mer probes) by utilizing the Stellaris® RNA FISH Probe Designer (Biosearch Technologies, Inc., Petaluma, CA) available online at www.biosearchtech.com/stellarisdesigner (version 4.2). The slides were hybridized with the *Shh* Stellaris FISH Probe set labelled with Quasar 570 (Biosearch Technologies, Inc.), following the manufacturer's instructions available online at www.biosearchtech.com/stellarisprotocols. All PBS and ddH₂O used during RNA FISH were treated with DEPC and autoclaved to inactivate RNase enzymes. RNase free consumables were used throughout and glassware treated with RNaseZAP (Invitrogen AM9780).

Paraffin embedded tissue sections were dewaxed and rehydrated by transferring through the following series; xylene, xylene (5 mins), 100% EtOH, 100% EtOH then 95% EtOH for ~10 mins each at room temperature unless stated otherwise. Slides were permeabilised in 70% ethanol for at least 24 hours (up to one week) at 4°C prior to hybridisation. Slides were washed in PBS, incubated in 10 µg/mL proteinase K in 1X PBS for 20 minutes at 37°C followed by 5 min washes in 1X PBS and wash buffer (2X SSC, 10% deionised formamide) at room temperature.

Shh RNA FISH probes were warmed to room temperature, diluted to 125 nM in Stellaris RNA FISH hybridisation buffer (#SMF-HB1-10) containing 10% formamide

and hybridised to slides overnight in humidified chamber at 37°C. Slides were washed twice for 30 minutes in wash buffer (2X SSC, 10% deionised formamide) at 37°C and rinsed in PBS. Slides were stained with 0.5 µg/mL DAPI, mounted using Vectashield and stored at 4°C prior to imaging.

2.8.2 RNA FISH of cell lines

Cells were seeded, fixed and permeabilised as described in 2.7.4. In some cases, cells were permabilised in 70% EtOH as described in 2.8.1. Slides were incubated in wash buffer (2X SSC, 10% deionised formamide) for 5 mins at room temperature then probe preparation, hybridisations, washes and mounting was carried out as described in 2.8.1.

2.9 Imaging

2.9.1 Image acquisition and deconvolution

Slides were imaged using a Photometrics Coolsnap HQ2 CCD camera and a Zeiss AxioImager A1 fluorescence microscope with a Plan Apochromat 100x 1.4NA objective, a Nikon Intensilight Mercury based light source (Nikon UK Ltd, Kingston-on-Thames, UK) and Chroma #89014ET (3 colour) single excitation and emission filters (Chroma Technology Corp., Rockingham, VT) with the excitation and emission filters installed in Prior motorised filter wheels. A piezoelectrically driven objective mount (PIFOC model P-721, Physik Instrumente GmbH & Co, Karlsruhe) was used to control movement in the z dimension. Step size for z stacks was set to 0.2 µm. Hardware control and image capture were performed using either the acquisition module of Volocity (PerkinElmer Inc, Waltham MA) or Nikon Nis-Elements software (Nikon UK Ltd, Kingston-on-Thames, UK). Images were deconvolved using a calculated PSF with the constrained iterative algorithm in Volocity (PerkinElmer Inc, Waltham MA).

2.9.2 DNA FISH spot measurements

DNA FISH spot measurement was carried out using the quantitation module of Volocity (PerkinElmer Inc, Waltham MA). Interprobe distance was calculated by segmenting the hybridisation signals, finding the centroid of each and then measuring

the distance between the centroids of each fosmid. Only alleles with single probe signals were analysed to eliminate the possibility of measuring sister chromatids. Approximately 100 loci (>50 nuclei) were measured for each condition and probe combination.

2.9.3 RNA FISH signal quantification

RNA FISH signal quantification was carried out using the quantitation module of Volocity (PerkinElmer Inc, Waltham MA). Expressing alleles were calculated by segmenting the hybridisation signals and scoring each nuclei as containing 0, 1 or 2 RNA signals. Approximately 100 loci (>50 nuclei) were measured for each condition.

2.10 Synthetic biology tools

2.10.1 TALE design and assembly

Transcription activator-like effector (TALE) DNA binding domains specific to the promoter of *Shh* and to the forebrain enhancer SBE2 had previously been designed and assembled by Nezha Benabdallah (Benabdallah et al., 2019). TALE DNA binding domains specific to the limb enhancer ZRS were designed using TAL Effector Nucleotide Targeter 2.0 software (Doyle et al., 2012) with the following options used: array min and max = 15, upstream base = T, no T or A at position 1, sites must end in a T and may not end in G/NN, base composition rules enforced and search reverse complement. The target DNA sequence of the TALEs used in this thesis are summarised in Table 2.12.

TALEs were assembled using a modified protocol (Benabdallah et al., 2019; Therizols et al., 2014). In brief, a DNA binding domain specific for a 15 nucleotide sequence was generated by the modular assembly of 4 pre-assembled multimeric TALE repeat modules (three 4-mer and one 3-mer) into a modified TALE backbone vector containing VP64-2A-eGFP (Figure 2.1). The backbone vector used for assembly of the ZRS TALE was provided by Yatendra Kumar and has been modified to replace the ampicillin resistance cassette with spectinomycin resistance. TALE modules were picked from glycerol stocks of module library plates (Addgene 1000000034),

incubated overnight at 37°C shaking at 225 rpm in 5 mL L-broth containing ampicillin (50 ng/μL) and DNA isolated using the QIAprep Spin Miniprep kit (Qiagen 27104) according to the manufacturer's instructions. Miniprep DNA was eluted in 50 μL elution buffer and quantified using the Qubit dsDNA broad range assay with the Qubit 4 fluorometer and stored at -20°C. TALE modules were assembled into backbone vector by setting up a 20 μL one-pot golden-gate reaction as follows: vector (100 ng), TALE modules (200ng each), 2 μL Tango buffer (ThermoFisher ER0451), 2 μL Esp3I (ThermoFisher ER0451), 2 μL T4 DNA ligase (New England Biosciences M0202M), 2 μL 2mM ATP (CHECK***) in ddH₂O. Golden-gate reaction was performed on a thermal cycler (for programme details see Table 2.10). Competent cells (Invitrogen LS18263012) were transformed with 5 μL of reaction according to the manufacturer's instructions and then half of the cells were plated onto agar containing spectinomycin (50 ng/μL).

Colonies were screened by PCR for fully assembled TALEs by first restreaking and then setting up a 30 μL reaction as follows: single colony, 15 μL DreamTaq Green PCR Master mix (2X; Thermo Scientific K1082) and 0.5 μM forward and reverse primers in ddH₂O. Colony PCR was performed on a thermal cycler (for programme details see Table 2.11), reaction products were run on a 1.2% agarose gel to identify positive colonies and these colonies were confirmed by Sanger sequencing (IGMM Technical Services). TALE-Δ constructs were made by removing the BamHI-BglII fragment containing VP64 from the fully assembled TALE-VP64 plasmid by restriction digest. All TALE-VP64 and TALE- Δ plasmids were miniprep using the QIAprep Spin Miniprep kit (Qiagen 27104) according to the manufacturer's instructions, eluted in 50 μL elution buffer, quantified using the Qubit dsDNA broad range assay with the Qubit 4 fluorometer and then stored at -20°C prior to transfection. Glycerol stocks of TALE plasmids were made by adding 500 μL of glycerol to 500 μL TALE culture and then stored at -80°C.

Table 2.10. Golden-gate assembly of TALEs thermal cycler programme.

Temperature (°C)	Time (minutes)	Cycles
37	10	10-12
16	10	
36	15	
80	5	
4	∞	

Table 2.11. Colony PCR of TALEs thermal cycler programme.

Temperature (°C)	Time (seconds)	Cycles
95	180	
95	30	30
55	30	
72	120	
4	∞	

Table 2.12 TALE target DNA sequence.

TALE	Target sequence
tShh	GCCTTTCGCACTATA
tSBE2	CCCTCCTCTATACCC
tZRS	CTTAATCCTATAGAT
tSBE6	CCCCTCTCACATTCC
tNE	ATTCAACTATAACCT

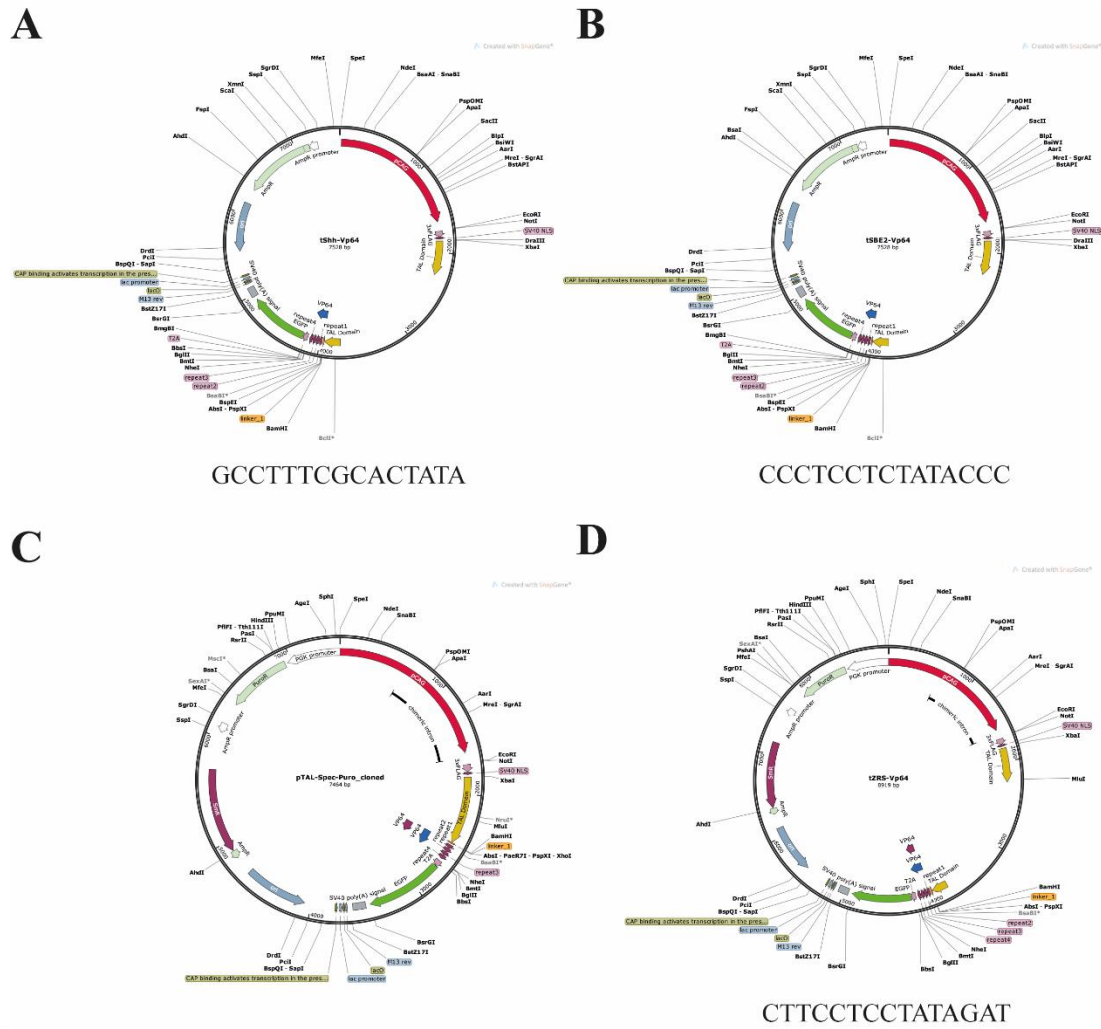


Figure 2.1. TALE construct maps.

Construct maps of TALE-VP64 plasmids with the target DNA sequence below for **(A)** tShh-VP64, **(B)** tSBE2-VP64, **(C)** modified TALE backbone vector provided by Yatendra Kumar and, **(D)** tZRS-VP64. tShh-VP64 and tSBE2-VP64 had previously been designed and assembled by Nezha Benabdallah (Benabdallah et al., 2019).

2.10.2 mESC transfections

mESCs were transfected with TALE plasmids using Lipfectamine 3000 Reagent (Invitrogen L3000015) according to the manufacturer's instructions. $2-3 \times 10^6$ mESCs were transfected with 14.5 µg of TALE plasmid and 26 µL Lipofectamine and seeded onto 0.1% gelatin coated 10 cm dish containing an autoclaved SuperFrost Plus Adhesion glass slide. Fresh media was added to cells after 24 hours. Slides were removed from dishes after 48 hours of transfection and washed, fixed and permeabilised for FISH as described in 2.7.4. The remaining cells in the dish were washed in PBS, treated with trypsin for ~2 minutes (mins) at 37°C to detach cells, centrifuged at 1200 rpm for 4 mins and cell pellets were either frozen at -80°C or used for RNA extraction.

2.10.3 Auxin-inducible degron induction

For induction of the auxin-inducible degron, indole-3-acetic acid (IAA, MP Biomedicals #102037) was added to the medium either 6 (Scc1-AID cells) or 24 (CTCF-AID cells) hours prior to cell collection. 500 µM of auxin (1000X IAA stock diluted in DMSO) was used for all depletion experiments and stored at 4°C for up to a month or at -20°C for long-term storage.

2.10.4 Expression analyses

RNA was extracted from cell pellets using the RNeasy mini kit (Qiagen 74106) as per the manufacturer's instruction, including an on-column DNase digestion (Qiagen 79254), eluted in 30 µL ddH₂O and quantified using the Qubit RNA broad range assay (ThermoFisher Q10210) with the Qubit 4 fluorometer. cDNA was synthesised from ~2 µg RNA using the Superscript II reverse transcriptase (Invitrogen 18064071) primed with random hexamers (Promega C1181) as per the manufacturer's instructions and stored at -20°C.

To quantify gene expression levels, a 20 µL qPCR was set up as follows: 10 µL Lightcycler 480 SYBR Green I Master (Roche 04707516001), 0.5 µM of forward and reverse primers and 1 µL cDNA in ddH₂O. A standard curve was include for each primer set analysed. qPCR was performed on a Lightcycler 480 (for programme details see Table 2.13).

Table 2.13. Gene expression analysis thermal cycler programme.

Temperature (°C)	Time (seconds)	Cycles
95	15 min hotstart	
95	30	44
55	30	
72	30	
4	∞	

2.11 Statistical tests

The Fisher's exact test is used to evaluate the significance of a non-random association occurring between two categorical variables, under the null hypothesis that the relative proportions of one variable is distinct from another second variable. I applied the Fisher's exact test to assess the difference between co-localised probe pairs between distinct conditions e.g. tissue types.

The Mann Whitney-U test is used to assess whether continuous variables from two independent samples (e.g. sample A and sample B) are significantly different from each other. This test does not presume a normal distribution of data and uses the null hypothesis that it is equally likely that a randomly selected value from sample A will be distinct from a randomly selected value from sample B. I applied the Mann Whitney-U test to compare quantitative variables (DNA FISH distances and RNA FISH quantification) between embryonic stages of development, different genotypes, tissue types and FISH probes.

Chapter 3. *Shh* regulation during neural development

3.1 Introduction

Shh expression is essential for embryonic brain and craniofacial development. *Shh*^{KO} mouse models lack midline and eye structures and have abnormal craniofacial features (Chiang et al., 1996). Heterozygous *SHH* mutations in humans is associated with HPE, a birth defect that occurs as a result of the forebrain failing to divide. HPE patients show a range of midline defects with the most severely affected patients displaying cyclopia whilst those more mildly affected can have craniofacial defects such as cleft palate/lip (Dubourg et al., 2007; Hehr et al., 2004; Roessler et al., 1996).

Shh neural expression is complex and many enhancers that regulate *Shh* expression in the brain have been identified including: forebrain-specific SBE2/3/4/7, midbrain-specific SBE1/5 and hindbrain-specific SFPE1/2 and SBE6 (Figure 1.9) (Benabdallah et al., 2016; Crane-Smith et al., 2019; Epstein et al., 1999; Jeong et al., 2006; Sagai et al., 2019; Yao et al., 2016). Many of these enhancers show overlapping regulatory domains and single enhancer knockouts do not severely disrupt embryonic development and have mostly subtle effects which is likely due to compensation from other redundant enhancers (Crane-Smith et al., 2019). Enhancer knockout models are limited as well as the difficulties with isolating enhancer-regulated cell populations from complex neural tissue makes it difficult to study *Shh* regulation in neural tissue *in vivo*. This means there is much unknown about the mechanism of these neural enhancers including the sequence of events related to their activation such as (1) when do these elements become accessible to transcription factor binding (2) what chromatin modifications are present and (3) which combination of trans-acting activators and repressors are required to ensure precise spatiotemporal enhancer activities?

In this chapter, I aimed to understand how the accessibility of the *Shh* regulatory landscape changes in *Shh*-expressing cells derived from fore-/midbrain tissues and which transcriptional regulatory networks are regulated by *Shh* during neural embryonic development by using ATAC-seq to identify nucleosome-free regions. I also investigated one of the *Shh* forebrain enhancers in more detail to understand its mechanism of long-range enhancer regulation and transcriptional regulation across different stages of development using DNA and RNA FISH.

3.2 ATAC-seq of *Shh*-expressing neural cells across development

3.2.1 Generation of ATAC-seq libraries from genetic mouse models

ATAC-seq (assay for transposable-accessible chromatin followed by sequencing) is a widely used method to find accessible chromatin that corresponds to *cis*-regulatory elements such as promoters and enhancers bound by proteins that exclude nucleosomes, including transcription factors (Buenrostro et al., 2013). The method utilises a hyperactive Tn5 transposase which ‘tags’ open chromatin with sequencing adaptors that can subsequently be amplified and sequenced (Figure 3.1A).

To look specifically at chromatin accessibility in *Shh*-expressing cells from neural tissue, I used embryos derived from *ShhGFP/cre* mice (hereafter referred to as *ShhGFP/cre*) or from *ShhGFP/cre* mice crossed to *mTmG* (*ShhGFP/cre;mTmG*) (for details see 2.4.1). I dissected forebrain/midbrain tissue from embryos by cutting between the maxillary and mandibular prominences across to the midbrain/hindbrain junction (Figure 3.1B). Two embryos were pooled together per experiment. Single-cell suspensions of dissected neural tissue were FAC-sorted into GFP⁺ (*Shh*-expressing) and GFP⁻ (non-expressing) cell populations. Cells were collected at three developmental time points for each genotype; E10.5, E11.5 and E14.5 (Figure 3.1C). *Shh*-expressing cells sorted from *ShhGFP/cre;mTmG* embryos represent descendants of a cell population expressing *Shh*, thereby acting as lineage markers of *Shh* expression during embryogenesis. *Shh*-expressing cells sorted from *ShhGFP/cre* embryos represent a cell population that are expressing *Shh* at that specific time point during embryogenesis. I obtained between 20,000-30,000 cells per sample following FACS and performed transposition on sorted cells (Figure 3.1D). For the remainder of this chapter, I will focus on the results from ATAC-seq data generated from *ShhGFP/cre* embryos.

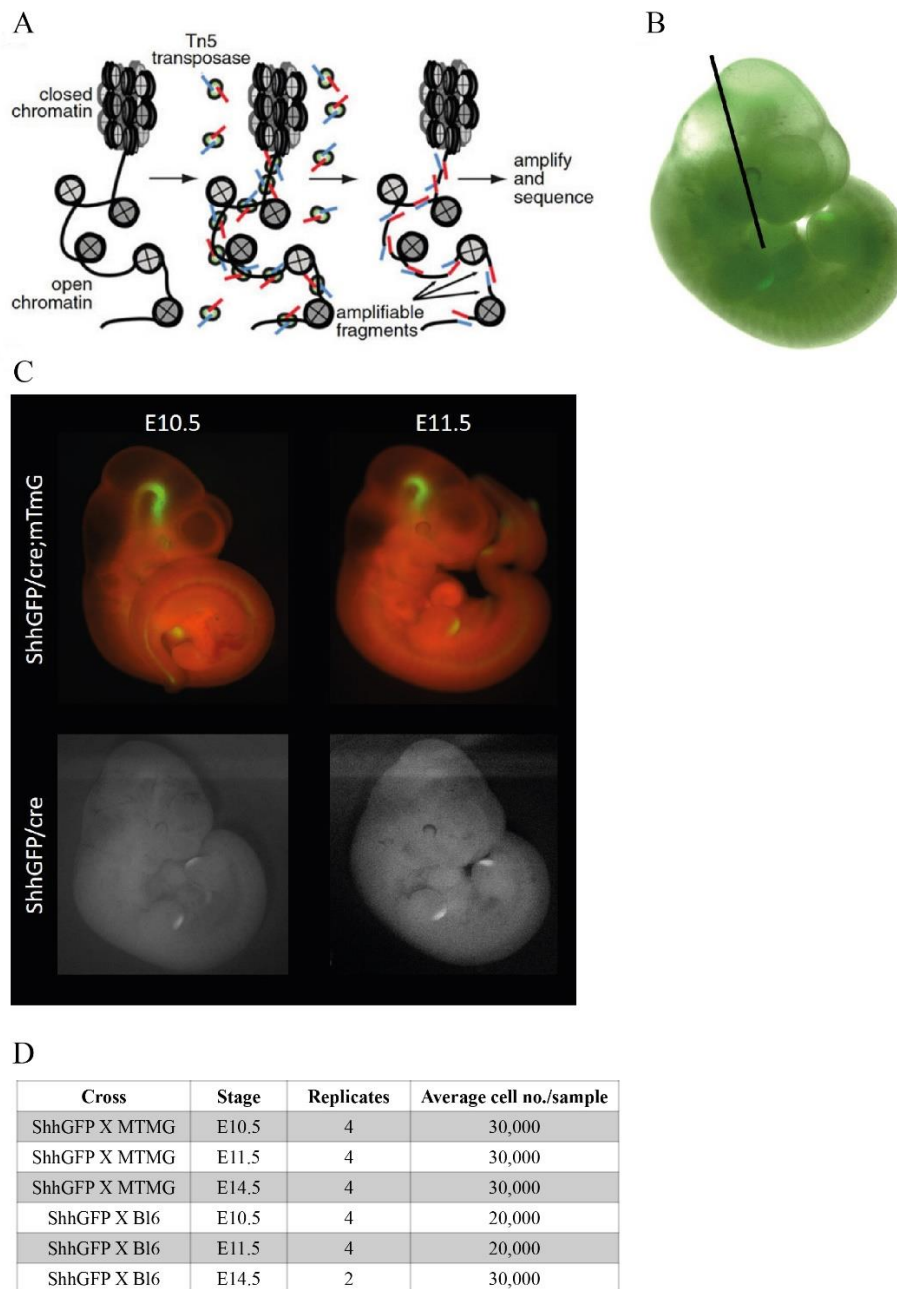
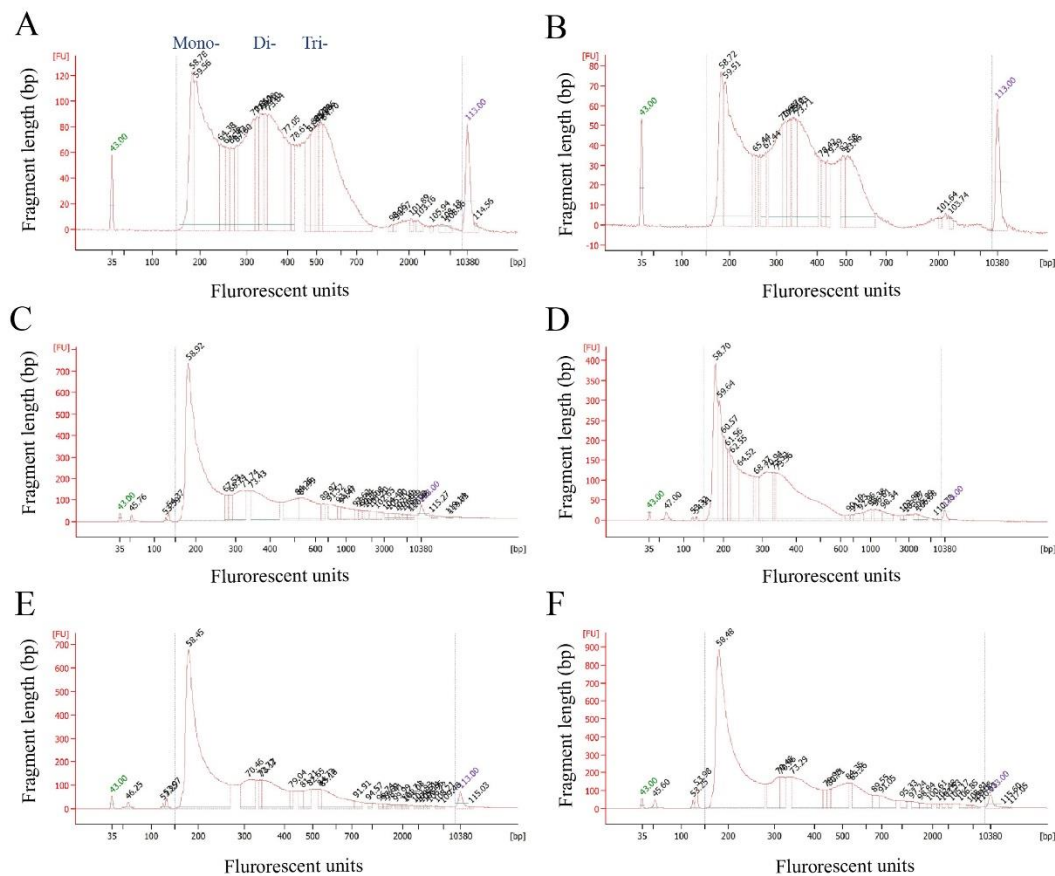


Figure 3.1 Isolation of *Shh*-expressing neural cells for ATAC-seq.

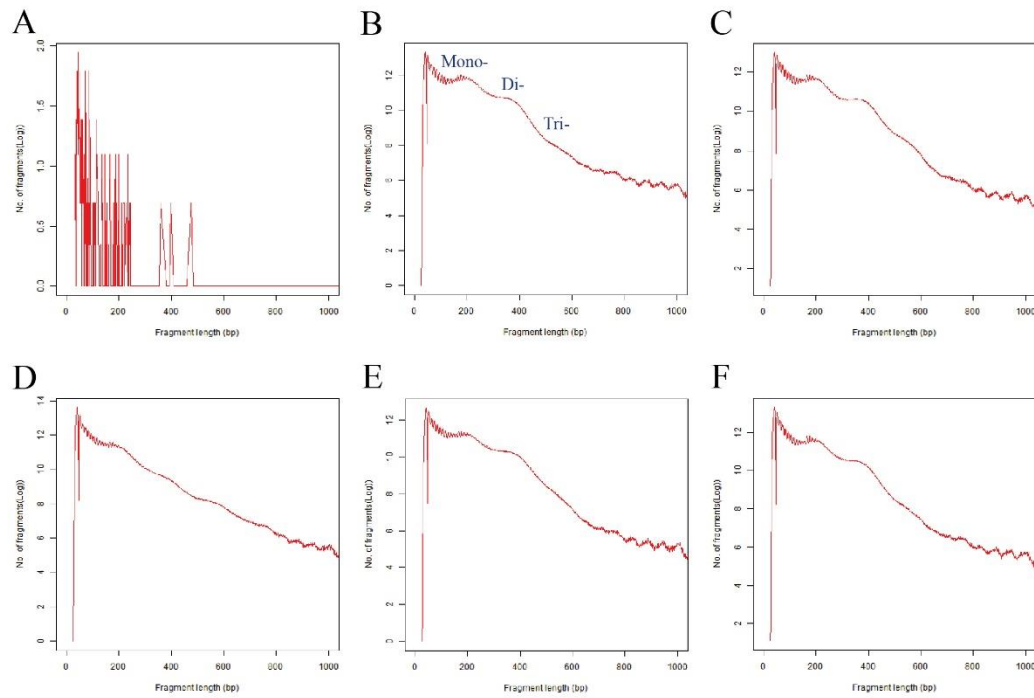
(A) Schematic of ATAC-seq method adapted from (Buenrostro et al., 2013). (B) Embryonic microdissections were performed between the maxillary and mandibular prominences across to the midbrain/hindbrain junction. Black line indicates approximate plane of dissection on E11.5 embryo. (C) Representative images taken from E10.5 and E11.5 ShhGFP/cre;mTmG and ShhGFP/cre embryos showing *Shh*-expressing cells. (D) Summary of samples obtained for ATAC-seq derived from ShhGFP/cre;mTmG and ShhGFP/cre dissected brains at E10.5, E11.5 and E14.5. The number of biological replicates and average number of *Shh*-expressing neural cells from each genotype/embryonic stage are detailed.



3.2.2 QC Analysis of ATAC-seq libraries

I checked the fragment length distribution of the insert size of sequenced fragments. Similar to the bioanalyser profiles, the majority of ATAC-seq samples had the expected distribution of mono-, di- and tri-nucleosomal peaks. Replicate 1 of GFP+ cells from an E10.5 *Shh*GFP/cre embryo showed an abnormal fragment length distribution (Figure 3.3A). Only 1901 reads were obtained for this samples suggesting this sample had not been sequenced (Figure 3.3G). This may be due to some error in library preparation. This sample was excluded from any further analysis. For all other samples from E10.5 *Shh*GFP/cre embryos, between 35 and 60 million reads were obtained for each sample and only a small proportion of mitochondrial reads were detected in aligned reads (~4-8%) (Figure 3.3G). All ATAC-seq samples obtained from E11.5 and E14.5 *Shh*GFP/cre embryos showed normal fragment length distributions (Appendix Figure 1 & Appendix Figure 2). Between 35 and 94 million reads were obtained per sample with a low proportion of mitochondrial reads (Appendix Figure 1 & Appendix Figure 2). Mitochondrial and duplicate reads were removed before further analysis.

Silvia Peluso (Hill Group) carried out principle component analysis (PCA) on ATAC-seq samples. For E10.5 *Shh*GFP/cre embryos very little amount of variation (0.2%) is detected between the samples (Figure 3.4A). Similarly, little variation is detected between samples from E11.5 (0.7%) or E14.5 (1.4%) *Shh*GFP/cre embryos (Figure 3.4B & C). When PCA was performed on all ATAC samples from *Shh*GFP/cre embryos, all samples cluster close together with only 0.9% variation found between these samples (Figure 3.4D). The samples did not cluster into *Shh*-expressing (GFP+) and non-expressing (GFP-) samples suggesting that overall the replicates are very similar even across different developmental stages.



G

E10.5 ShhGFP X Bl6	GFP+			GFP-		
Replicate	1	2	3	1	2	3
Read-pairs before clipping:	1901	59878608	51814350	55650114	34583968	54434538
Read-pairs after clipping:	1817	58036148	49817089	53310720	33178849	52292058
all read-pairs aligned:	341	56164122	47914059	50662297	31641758	49584959
Mitochondrial read-pairs in aligned:	44	3685293	2365345	3968626	1385542	3481689
Nuclear read-pairs after deduplication:	327	44328482	38691693	36574745	26462946	39318686
% chrM	12.90323	6.56165	4.936641	7.83349	4.37884	7.021664
% aligned	18.7672	96.77438	96.17997	95.0321	95.36726	94.82312
% nuclear	17.20147	74.03058	74.6737	65.72268	76.51796	72.23114

Figure 3.3 QC analysis of E10.5 ATAC samples following sequencing.

(A-F) Fragment length distribution profiles following sequencing of libraries prepared from E10.5 isolated neural cells from GFP+ replicate (A) 1, (B) 2, and (C) 3 and GFP- replicate (D) 1, (E) 2 and (F) 3. Peaks are shown for mono-, di- and tri-nucleosomal fragments (bp). (G) Summary of the read-pairs obtained for each ATAC-seq sample including the number of read-pairs before and after clipping and the number and percentage of mitochondrial, aligned and nuclear read-pairs after removal of duplicates.

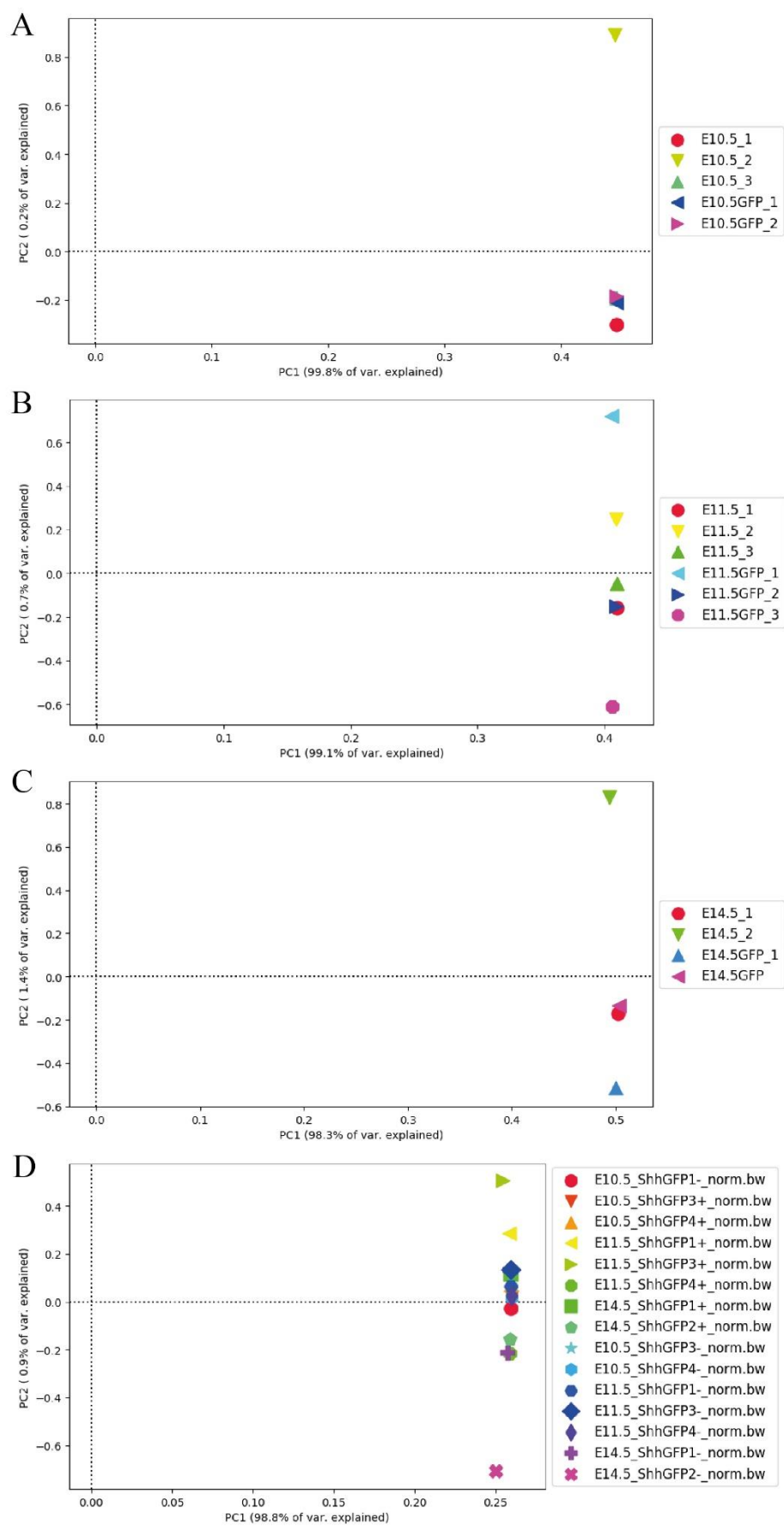
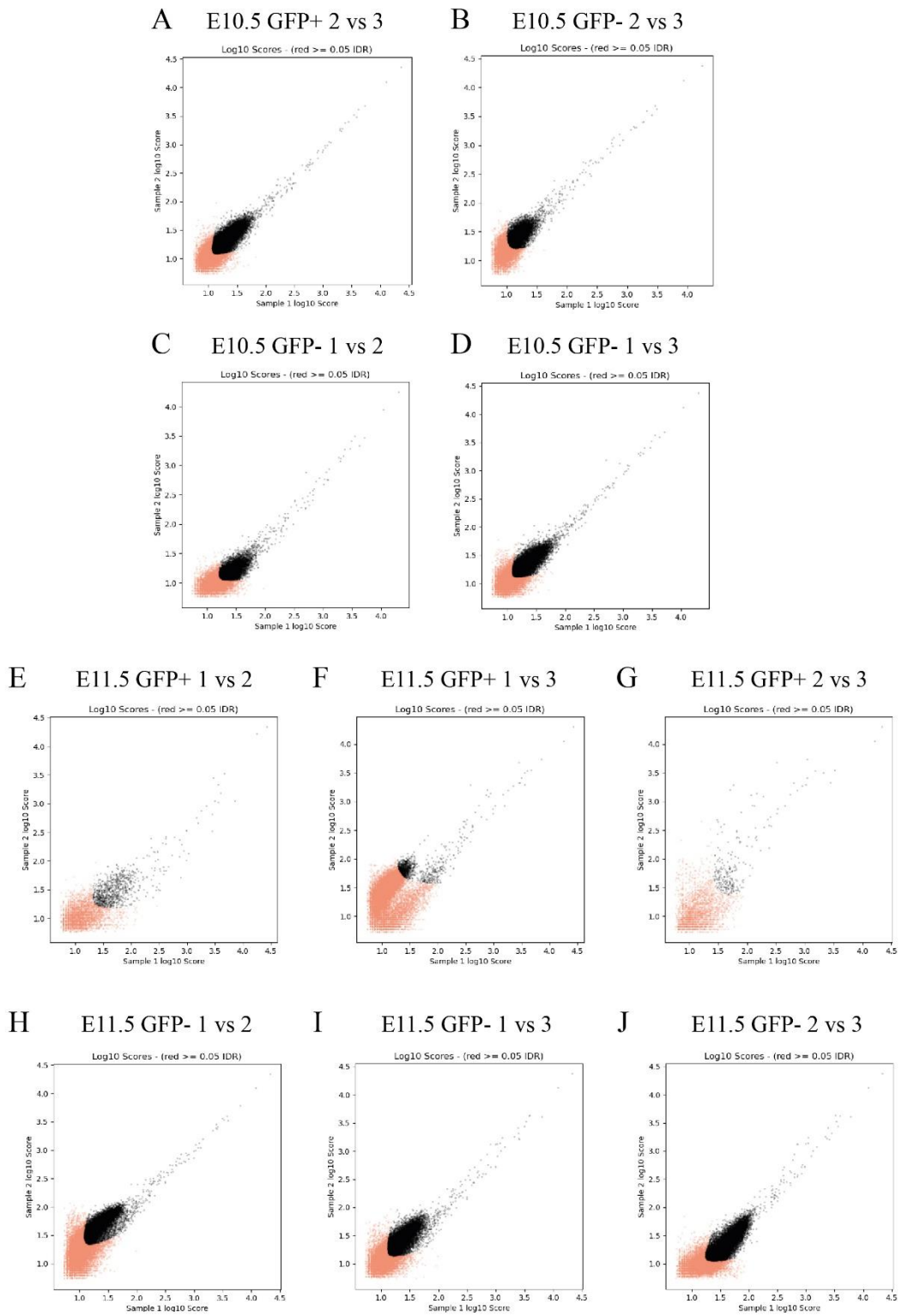
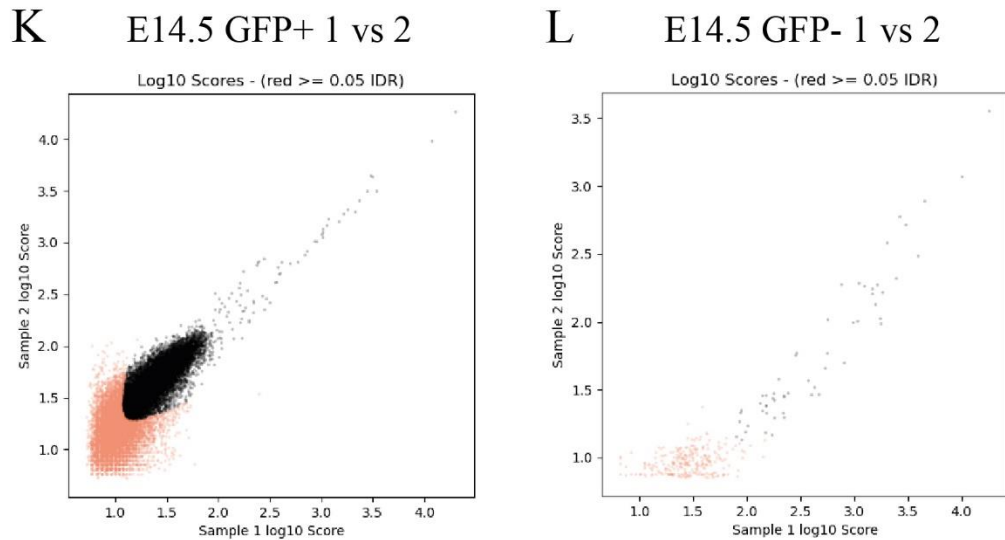


Figure 3.4 PCA analysis of ATAC samples.

(A-C) Principle component analysis (PCA) of ATAC samples following sequencing of libraries prepared from (A) E10.5, (B) E11.5 and (C) E14.5 GFP+ or GFP- isolated neural cells. Biological replicates are numbered (e.g. E10.5_1). GFP+ samples are indicated as e.g. E10.5GFP_1. (D) PCA of all ATAC samples. This analysis was carried out by Silvia Peluso.

To assess the consistency between replicates, I implemented Irreproducibility Discovery Rate (IDR) analysis. IDR is a statistical method that compares the peaks called between a pair of replicates and scores the reproducibility between them. The premise of IDR analysis is that the most significantly peaks called are likely true signal and would be expected to be highly consistent between replicates whereas less significant peaks are likely noise and will not be consistent between replicates. Here, I called narrow peaks using MACS2 and performed IDR using an IDR cutoff of 0.05 meaning that a peak has a 5% chance of being an irreproducible discovery. Each peak is scored between 0 and 1. The peak scores are plotted as log10 values and any peaks that do not pass the IDR threshold are coloured red. For E10.5 ShhGFP/cre samples, approximately 50-65% of called peaks were reproducible between the replicates (Figure 3.5A-D & M). Low reproducibility is found between the replicates of E11.5 GFP+ samples with only 9-23% of called peaks passing the IDR threshold (Figure 3.5E-G & M). In comparison, higher reproducibility is observed between the replicates of E11.5 GFP- samples with approximately 50-64% of called peaks passing the IDR threshold (Figure 3.5H-J & M). Approximately 60% of called peaks were reproducible between the replicates of E14.5 GFP+ samples whereas only 17% passed the IDR threshold for E14.5 GFP- replicate samples (Figure 3.5K-M). It is unclear why there is low reproducibility between the E11.5 GFP+ replicate samples and this experiment will need to be replicated.





M

<u>Sample</u>	<u>Replicates</u>	<u>Common peaks</u>	<u>>0.05</u>	<u>Peaks >0.05 (%)</u>
E10.5 GFP+	2 vs 3	15589	10194	65.4
E10.5 GFP-	1 vs 2	11013	5662	51.4
E10.5 GFP-	2 vs 3	10939	6083	55.6
E10.5 GFP-	1 vs 3	15377	9133	59.4
E11.5 GFP+	1 vs 2	3900	893	22.9
E11.5 GFP+	2 vs 3	2576	255	9.9
E11.5 GFP+	1 vs 3	13328	1179	8.8
E11.5 GFP-	1 vs 2	18689	9468	50.7
E11.5 GFP-	2 vs 3	17652	11283	63.9
E11.5 GFP-	1 vs 3	15547	8626	55.5
E14.5 GFP+	1 vs 2	18536	11048	59.6
E14.5 GFP-	1 vs 2	321	56	17.4

Figure 3.5 IDR analysis of ATAC samples.

Irreproducible discovery rate (IDR) analysis of ATAC peaks called from E10.5 GFP+ (A), E10.5 GFP- (B-D), E11.5 GFP+ (E-G), E11.5 GFP- (H-J), E14.5 GFP+ (K) and E14.5 GFP- (L) replicate samples. IDR scores are shown as log10 values. Peaks that pass the IDR threshold of 0.05 are black and peaks that fail are red. (M) Table summarising the common peaks found between replicate ATAC samples and the number and percentage of peaks scoring above the IDR threshold of 0.05.

3.2.3 Chromatin accessibility in neural tissue is unaffected by *Shh* of developmental stage

Silvia Peluso (Hill Group) performed both broad and narrow peak calling on merged ATAC-seq samples using MACS2 (<https://macs3-project.github.io/MACS/>) (Zhang et al., 2008). It should be noted that this analysis was performed independently and did not take into account the findings of the IDR analysis. In the future only common peaks that are reproducible between replicates should be used for further analysis. As I was interested in identifying transcription factor binding sites in the ATAC-seq data, I focused on the narrow peaks and generated bigWig files for visualisation of each merged dataset on UCSC genome browser (<https://genome.ucsc.edu>).

3.2.3.1 ATAC-seq profile of the *Shh* regulatory landscape

I first looked at the ATAC-seq profile of the *Shh* regulatory landscape at each developmental time point from both *Shh*-expressing and non-expressing cells. Lower ATAC signal is detected from non-expressing E14.5 cells compared to the rest (Figure 3.6). The previous IDR analysis showed there was little reproducibility between the E14.5 GFP- samples (Figure 3.5L & M) Viewing the entire *Shh* TAD, I found that the overall pattern of accessible chromatin was similar across the stages of development. The most prominent ATAC signal is detected at the promoters of *Lmbr1* and *Rnf32*, two ubiquitously expressed genes (Figure 3.6). There is also increased accessibility across the *Shh* gene body in both *Shh*-expressing and non-expressing cells regardless of developmental stage (Figure 3.6).

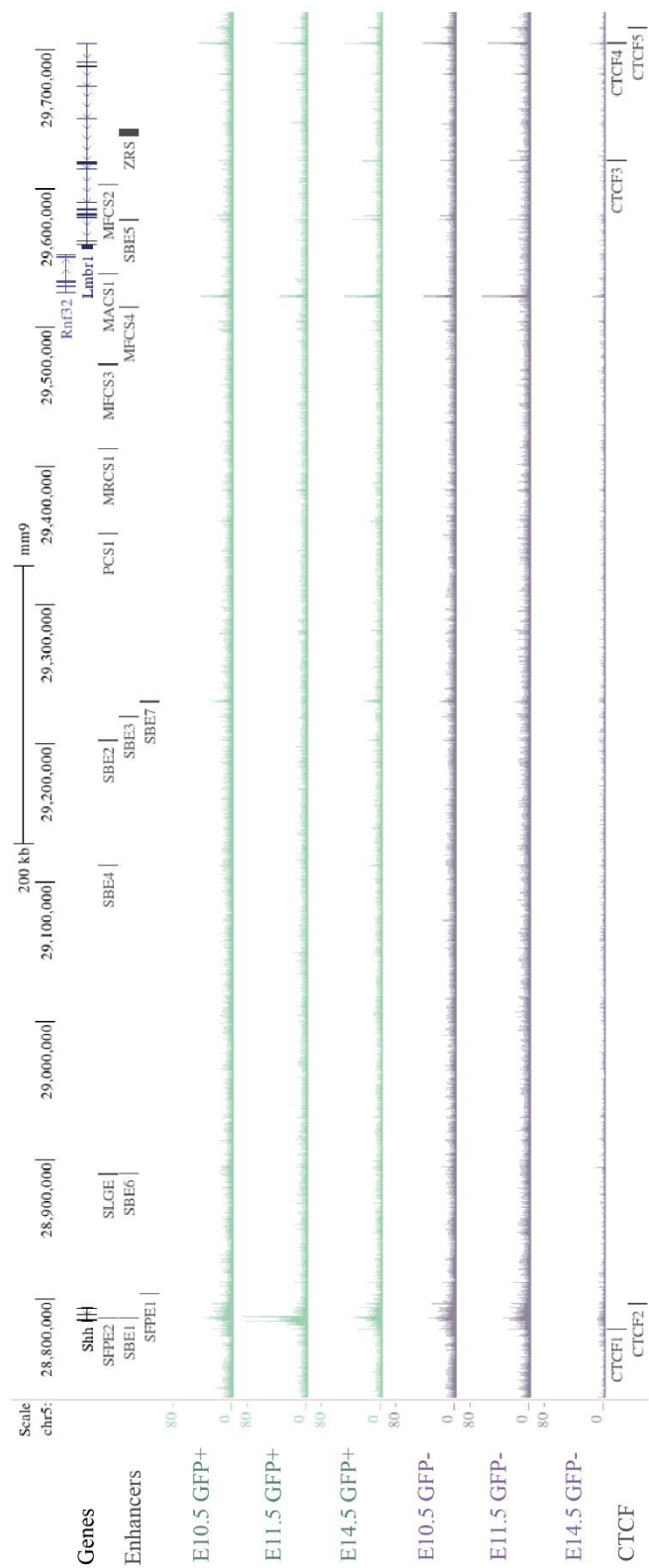


Figure 3.6 Chromatin accessibility across the *Shh* regulatory landscape.

Genome browser view of the *Shh* TAD (mm9 chr5: 28,728,517-29,727,005) showing the location of genes and the known *Shh* enhancers at the top and the position of the CTCF sites found at either end of the TAD are numbered below (1-5). Tracks show ATAC-seq profile from *Shh*-expressing (GFP+, green) or non-expressing (GFP-, purple) at E10.5, E11.5 and E11.5 cell populations derived from embryonic fore-/midbrain tissue.

Upon closer inspection of the region surrounding the *Shh* gene up to the neural enhancer SBE6, I detected chromatin accessibility across the midbrain enhancer SBE1 (black arrow) (Figure 3.7A). The ATAC profile across the *Shh* gene body from E11.5 GFP+ cell population looked different to the others with a large peak found adjacent to SBE1 and another found at the end of the gene (Figure 3.7A). It was unclear why the profile looks different here but as shown previously there is some inconsistency between these replicates (Figure 3.5E-G). The chromatin underlying the floorplate enhancers SFPE1 and SBE6 did not show accessibility (Figure 3.7A). It was hard to discern if SFPE2 is accessible due to it being located within the *Shh* gene body and in close proximity to SBE1. Chromatin accessibility peaks could also be detected at CTCF sites 1 and 2 (black asterisk) (Figure 3.7A). The centre of the *Shh* TAD that encompasses the forebrain enhancers; SBE2, SBE3, SBE4 and SBE7. SBE7 is accessible at E10.5 and E14.5 in *Shh*-expressing and at E10.5 and E11.5 in non-expressing cells (black arrow) (Figure 3.7B). No ATAC-signal is detected for SBE3 or SBE4. Finally, I looked more closely at a region at the other end of the *Shh* TAD that includes *Lmbr1*. With the exception of E11.5 GFP+ ATAC profile, I found that the midbrain enhancer SBE5 was accessible in both *Shh*-expressing and non-expressing cells across development (black arrow) (Figure 3.7C). ATAC-seq signal could also be detected at CTCF sites 3, 4 and 5. A novel region of chromatin accessibility is found in intron 3 of *Lmbr1* in both GFP+ and - cells (Figure 3.7).

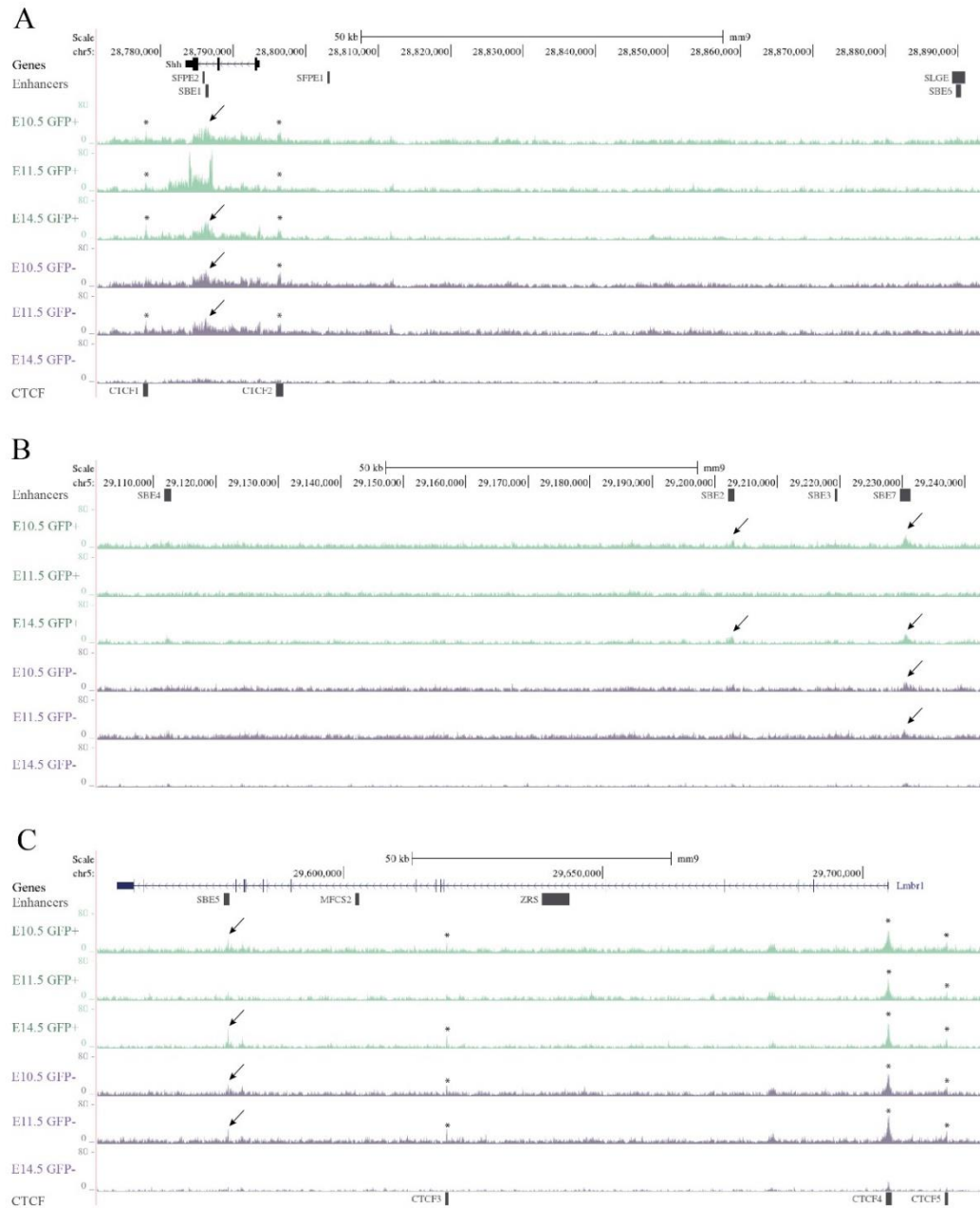


Figure 3.7 *Shh* neural enhancers and CTCF sites can be detected by ATAC-seq.

(A-C) Genome browser views of the *Shh* TAD from (A) *Shh* to SBE6 (chr5: 28,771,242-28,893,861), (B) SBE4 to SBE7 (chr5: 29,100,985-29,243,410) and (C) across *Lmbr1* (chr5: 29,552,650-29,723,711) showing the location of genes and the known *Shh* enhancers at the top and the position of the CTCF sites at found at either end of the TAD are numbered below (1-5). Tracks show ATAC-seq profile from *Shh*-expressing (GFP+, green) or non-expressing (GFP-, purple) at E10.5, E11.5 and E11.5 cell populations derived from embryonic fore-/midbrain tissue. Chromatin accessibility peaks detected at enhancers are indicated by arrows and peaks corresponding to CTCF sites are shown by an asterisk.

3.2.3.2 ATAC-seq profiles of other regulatory landscapes

Overall I found little difference in the chromatin accessibility landscape across the development stages and between *Shh*-expressing vs non-expressing cells. To see if this observation was limited to the *Shh* regulatory landscape I looked at the ATAC-seq profile of other genes known to be involved in embryonic head development. *Otx2* encodes a transcription factor with roles in early anterior neural plate patterning and later maintains forebrain and midbrain identities (Simeone et al., 2003). I found that regions containing long-range anterior neuroectoderm (AN1/2) and forebrain/midbrain (FM1) enhancers were accessible across all developmental stages (Figure 3.8A). No accessible regions were detected for any of the 3' long-range *Otx2* enhancers (Chan et al., 2020). The chromatin region upstream of the *Otx2* also showed ATAC-seq signal and likely corresponds to proximal enhancers (Figure 3.8A). A novel accessibility peak was detected 58 kb upstream of *Otx2* (red arrow) (Figure 3.8A). I also detected ATAC-seq signal at a region that seems to correspond to a CTCF binding site (asterix) (Figure 3.8A).

The midbrain-hindbrain boundary is defined by the antagonising activities between *Otx2* and *Gbx2* – a transcription factor that defines hindbrain identity (Simeone et al., 2003). I found that an ATAC-seq signal could be detected at the *Gbx2* promoter and its anterior most-hindbrain 1 (AMH1) enhancer (Figure 3.8B). As a control, I looked at the regulatory landscape of the α -globin genes which are normally expressed exclusively in erythroid cells under the control of a cluster of enhancers (De Gobbi et al., 2007; Hughes et al., 2005). No chromatin accessibility peak is detected at any of the α -globin enhancers in any of the cell populations (arrowheads) (Figure 3.8C). Accessibility is detected over gene promoters (*Mpg* and *Nprl3*) and CTCF sites only (asterix) (Figure 3.8C).

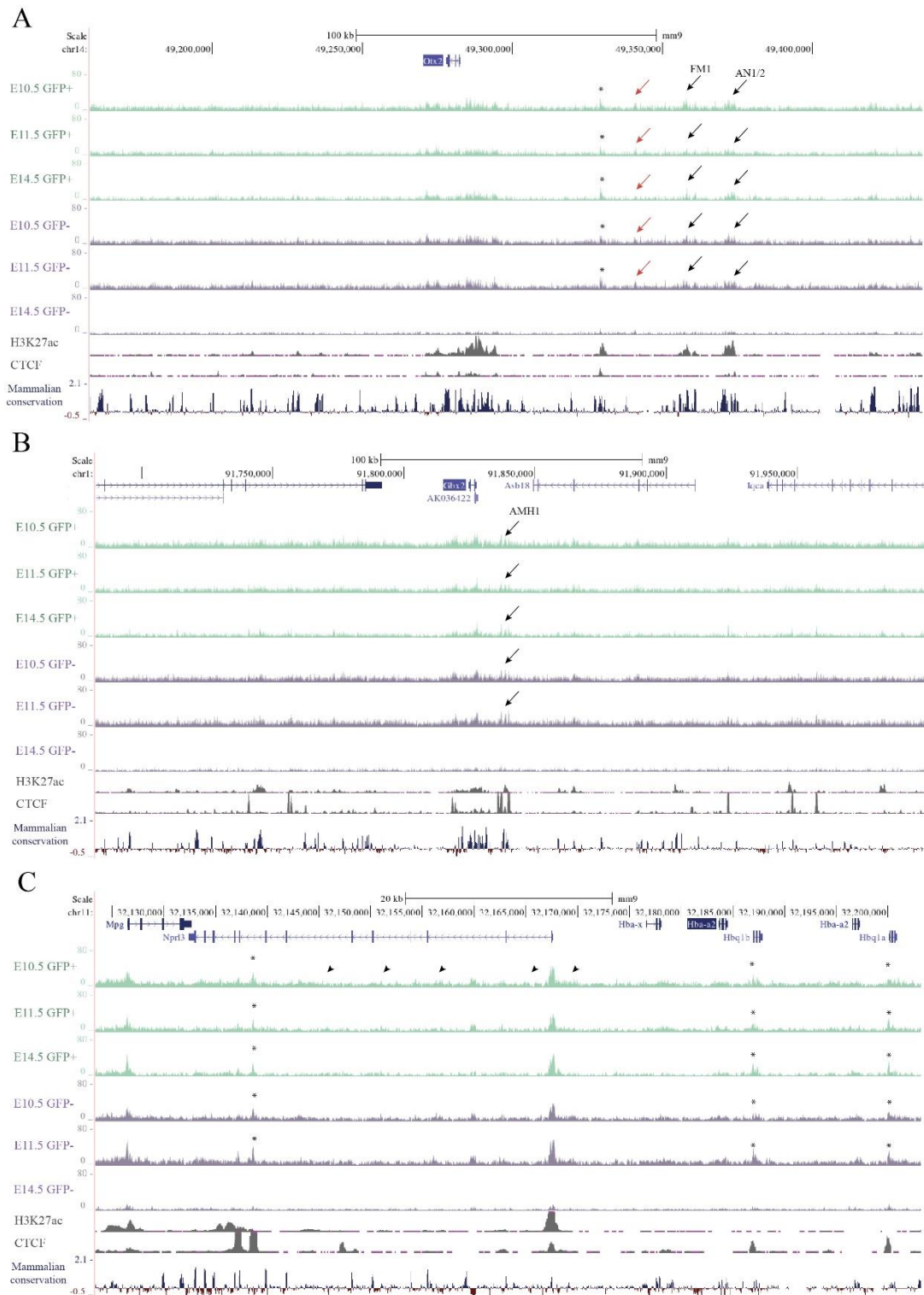


Figure 3.8 Chromatin accessibility profiles of other gene regulatory landscapes.

(A-C) Genome browser views of the regulatory landscapes of the (A) *Otx2*, (B) *Gbx2* and (C) α -globin genes. The location of genes are shown at the top and the bottom

tracks show UCSC ENCODE H3K27ac and CTCF ChIP-seq profiles from E14.5 brain tissue and mammalian conservation. Central tracks show ATAC-seq profile from *Shh*-expressing (GFP+, green) or non-expressing (GFP-, purple) at E10.5, E11.5 and E14.5 cell populations derived from embryonic fore-/midbrain tissue. Chromatin accessibility peaks detected at known (black) or novel (red) enhancers are indicated by arrows, peaks corresponding to CTCF sites are shown by an asterisk and the location of the α -globin enhancers are indicated by arrowheads.

3.2.4 Analysis of peaks unique to *Shh*-expressing cell populations

Little overall difference in the chromatin accessibility was observed between *Shh*-expressing and non-expressing cell populations when looking at all peaks called from each ATAC-seq sample. Silvia Peluso (Hill Group) overlapped the merged narrow peak files from *Shh*-expressing (GFP+) and non-expressing (GFP-) at each development stage. She then extracted the peaks that were unique either to *Shh*-expressing (GFP+) and annotated the genomic distribution of the GFP+ unique peaks (Figure 3.9). It should be noted that this analysis was performed independently and did not take into account the findings of the IDR analysis. In the future only common peaks that are reproducible between replicates should be used for this type analysis. I used the GFP+ unique peaks to perform gene ontology (GO) analysis of biological processes for each developmental stage and looked at the top 10 most significant GO terms.

The majority of peaks unique to E10.5 *Shh*-expressing cells were found in either intronic (37%) or intergenic (36%) genomic regions suggesting that the majority of chromatin accessibility are within regulatory regions (Figure 3.10A). The majority of GO terms identified here are processes such as cell differentiation and general development (Figure 3.10A). Nervous system development and its regulation are also significant biological processes which is unsurprising as only neural tissue is used in the ATAC-seq experiments (Figure 3.10A). There is an increased proportion of unique peaks found in intergenic (54%) regions from E11.5 *Shh*-expressing cells indicating that more regulatory regions are becoming accessible at this stage of development (Figure 3.10B). Again, many of the significant GO terms found are related to cell and/or nervous system development and differentiation as well as regulation of these processes. Unique peaks from E11.5 *Shh*-expressing cell appear to also be linked to

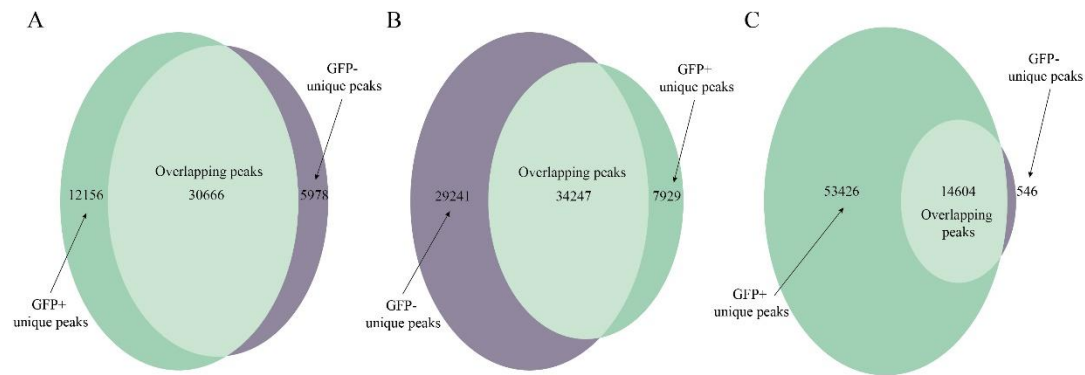


Figure 3.9 Identification of unique accessible regions in *Shh*-expressing cells.

(A-C) Venn diagrams showing the number of peaks that overlap (light green) or are unique to either *Shh*-expressing (GFP+, green) or non-expressing (GFP-, purple) cells derived from (A) E10.5, (B) E11.5 or (C) E14.5 fore-/midbrain tissue. This analysis was carried out by Silvia Peluso.

processes involving neuron differentiation, development of neuron projections and Wnt signalling regulation (Figure 3.10B). Similarly to E10.5, the majority of peaks unique to E14.5 *Shh*-expressing cells were found in either intronic (37%) or intergenic (35%) regions (Figure 3.10C). GO term analysis from this later stage of development of *Shh*-expressing cells indicates that biological processes that regulate axon generation and extension are being favoured which had previously not been found at the earlier stages (Figure 3.10C).

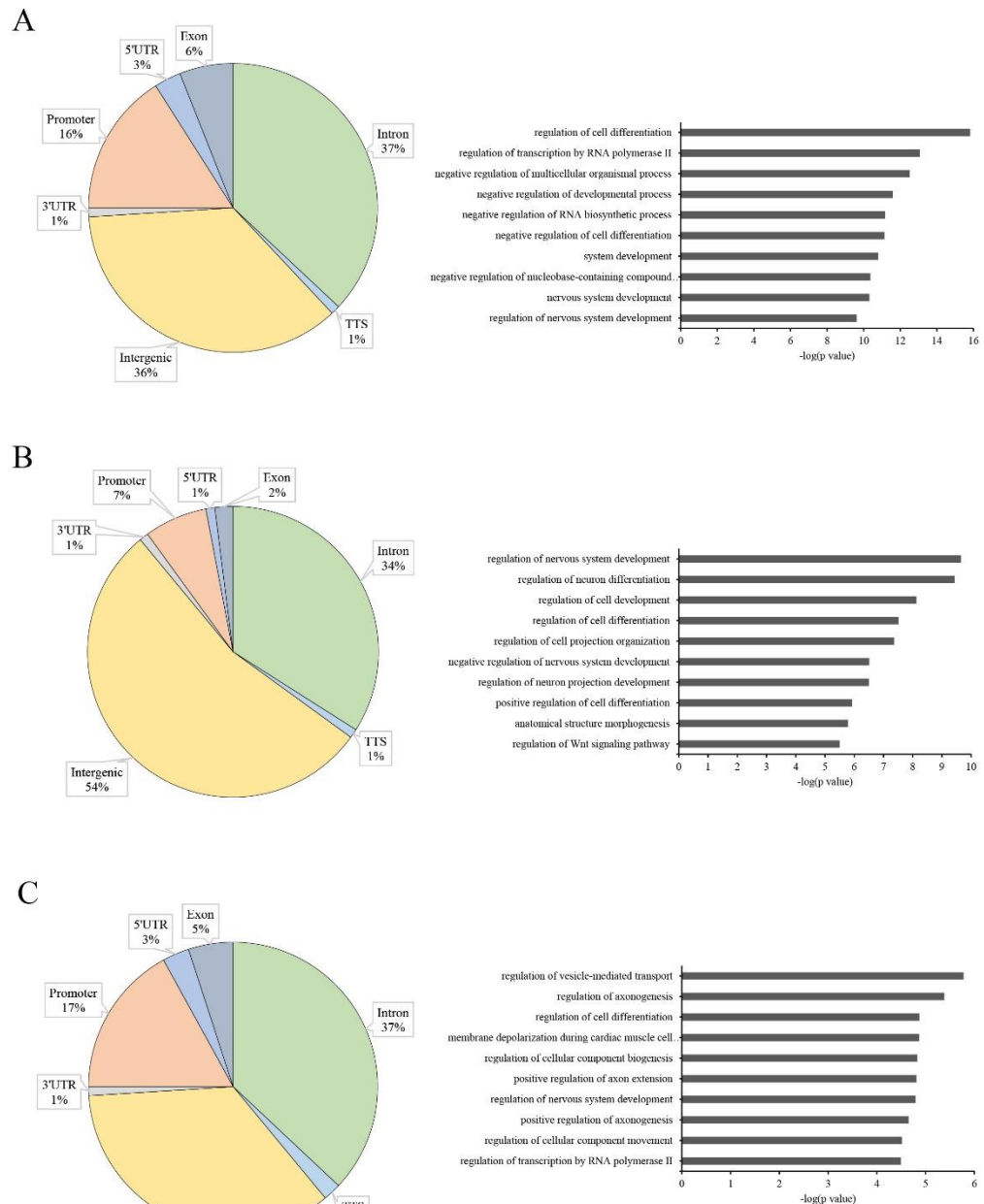


Figure 3.10 Genomic distribution and gene ontology analysis of unique peaks from *Shh*-expressing cells.

(A-C) Unique peaks identified in *Shh*-expressing cells from (A) E10.5, (B) E11.5 and (C) E14.5 fore-/midbrain tissue were used to calculate the genomic distribution of unique accessible regions (left hand panels) and to perform gene ontology analysis to identify highly significant ($-\log(p \text{ value})$) biological processes associated with these stage-specific unique peaks (right hand panels). Left hand graphs were generated by Silvia Peluso.

3.2.5 Motif analysis across developmental stages

To try to understand which transcriptional regulatory processes are occurring in *Shh*-expressing tissue in comparison to non-expressing neural tissue across the different developmental stages I decided to investigate which motifs are found enriched within the ATAC-seq peaks. To identify enriched motifs within the ATAC-seq samples, Silvia Peluso (Hill Group) performed motif analysis using HOMER (<http://homer.ucsd.edu/homer/>) (Heinz et al., 2010). From the HOMER *de novo* motif output I extracted the top 10 enriched motifs and any other motifs that were found in >5% of peaks (Appendix Figure 3, Appendix Figure 4 & Appendix Figure 5). I focused on these motifs for further analysis.

3.2.5.1 Commonly occurring motifs

One of the most highly ranked and common motif detected across all samples was CTCF (Figure 3.11). This was unsurprising as CTCF binding is found in every cell/tissue type and I showed that I could detect chromatin accessibility that corresponded to CTCF binding at the *Shh*, *Otx2* and *Gbx2* regulatory landscapes (Figure 3.7 Figure 3.8). A highly enriched motif predicted to be bound by the KLF/SP family of transcription factors was common to all ATAC-seq samples (Figure 3.11). Several proteins from the subfamilies of the ETS transcription factor family including TCF, ERG, PEA3, SPI, ELG, TEL and ETS were predicted to bind a motif found in all ATAC-seq samples (Figure 3.11). KLF/SP and ETS transcription factors can function as both transcriptional activators and repressors and are involved in the regulation of a wide range of developmental processes and are not specific to neurogenesis (Bialkowska et al., 2017; Maroulakou and Bowe, 2000).

Motifs that are predicted to be bound by RFX or SOX transcription factors are more highly enriched in E10.5 *Shh*-expressing cells, then in non-expressing E11.5 and become prevalent again in *Shh*-expressing cells at E14.5 (Figure 3.11). Motifs predicted to be bound by RFX proteins are also found in E10.5 non-expressing cells. RFX proteins have functions in ciliogenesis and *Rfx4* in particular is required for development of midline structures of the fore- and midbrain (Zhang et al., 2007). The SOX family of transcription factors are important for neural cell development and act

sequentially during the differentiation process to specify neurons (Bergsland et al., 2011). The CAAT box motif was another highly ranked motif found across all ATAC-seq samples with the exception of E14.5 *Shh*-expressing cells (Figure 3.11). NFI and MSX are predicted to bind this motif and these transcription factors are known to function in telencephalon (Mason et al., 2009) and craniofacial development (Alappat et al., 2003), respectively, unrelated to *Shh* expression.

3.2.5.2 Motifs exclusive to cell type or developmental stage

The motifs highlighted thus far are found recurrently across the different developmental stages irrespective of *Shh* expression. I next looked more closely at motifs that were unique to either *Shh*-expressing cells, a developmental stage or a particular cell population. A motif predicted to be bound by transcription factors such as CRX, GSC, DMBX1 showed some differences between *Shh*-expressing and non-expressing cells at E10.5 and E11.5 stages of development (Figure 3.11). At E10.5 this motif is predicted to be bound by retinoic acid receptors (RAR) in *Shh*-expressing cells (Figure 3.11). In non-expressing at E10.5 and E11.5 and in both cell populations from E14.5 this motif is not predicted to be bound by RAR but instead to be bound by OTX2 and PITX transcription factors (Figure 3.11). Both of these transcription factors have roles in establishing the anterior-posterior axis in the developing embryo (Arenas et al., 2015; Simeone et al., 2003).

As shown previously, the ATAC-seq peaks from *Shh*-expressing from E11.5 staged embryos seem to be involved in biological processes such as regulation of neuronal differentiation and WNT signalling. I identified a motif present only in E11.5 *Shh*-expressing cell populations that could potentially explain a link between SHH, WNT and neuronal differentiation, TCF3 (Figure 3.11). In the neural tube, dorsal-ventral patterning is controlled by the antagonising activities of SHH and WNT signalling (Alvarez-Medina et al., 2008). *Tcf3* is highly expressed at E11.5 and down-regulated from E13.5 in neural stem cells and is a regulator of neural stem cell identity by repressing the WNT signalling pathway (Kuwahara et al., 2014; Ohtsuka et al., 2011). In *Shh*-expressing cells at E11.5, TCF3 may also be functioning to maintain neural stem cell identity by binding to and regulating factors involved in WNT signalling.

Transcription factor family	Example binding motif	Cell population motif detected in
CTCF		All populations
KLF SP		All populations
ETS		All populations
RFX		E10.5 <i>Shh</i> -expressing E10.5 non-expressing E11.5 non-expressing E14.5 <i>Shh</i> -expressing
SOX		E10.5 <i>Shh</i> -expressing E11.5 non-expressing E14.5 <i>Shh</i> -expressing
NFI MSX		E10.5 <i>Shh</i> -expressing E10.5 non-expressing E11.5 <i>Shh</i> -expressing E11.5 non-expressing E14.5 non-expressing
RAR CRX GSC DMBX		E10.5 <i>Shh</i> -expressing
OTX2 PITX CRX GSC DMBX		E10.5 non-expressing E11.5 non-expressing E14.5 <i>Shh</i> -expressing E14.5 non-expressing
TCF3		E11.5 <i>Shh</i> -expressing

Figure 3.11 Motif analysis of across neural development

Summary of the commonly occurring and exclusive motifs identified showing the associated transcription factor family, a representative profile of the binding motif and the cell populations associated with a particular motif.

3.3 *Shh* forebrain expression is distally regulated

We have previously shown using both 3C-based and fluorescence *in situ* hybridisation (FISH) methods that *Shh* and ZRS are found in close proximity across all cell types and tissues analysed thus far with the overall TAD structure being largely invariant (Williamson et al., 2016, 2019). Increased co-localisation is observed between *Shh* and ZRS in posterior limb bud cells with the intervening region further away from both these loci, consistent with an enhancer-promoter loop forming between *Shh* and ZRS (Williamson et al., 2016). As discussed in 1.3.5, *Shh* has a complex regulatory landscape and it is unclear how the other distal *Shh* enhancers operate within this apparent invariant TAD structure.

To address this question, I decided to focus on the forebrain enhancer, SBE2. Previous work using an enhancer trap assay to screen across the *Shh* regulatory region identified multiple distal forebrain enhancers, including SBE2 which is located approximately 400 kb upstream of *Shh* in the centre of the TAD (Jeong et al., 2006). SBE2 regulates *Shh* expression in the rostral diencephalic division of the forebrain at ~E8-E12 embryonic stages of development. Point mutations in this enhancer disrupts normal brain and craniofacial development and has been implicated in HPE (Jeong et al., 2008b). SBE2 knockout mouse models lose *Shh* expression in the rostral diencephalon have mild deformities with dorsal-ventral patterning defects (Crane-Smith et al., 2019). DNA FISH carried out by Iain Williamson (Bickmore Group) in mESCs and in the developing limb found that SBE2 is ~400 nm away from *Shh* (Williamson et al., 2019). This data is consistent with an invariant *Shh* TAD structure with the intervening region (including the forebrain enhancers SBE2, 3 and 4) being looping out. Others in our lab have shown that one of the neural enhancers, SBE6, moves further away from *Shh* upon its activation (Benabdallah et al., 2019). However, SBE2 is not active in the cells analysed in any of these prior studies and it remains unclear what the *Shh* structure may look like in forebrain tissue where SBE2 is regulating *Shh* expression. I did not find chromatin accessibility at SBE2 in the ATAC-seq data but this may be due to the majority of the cells coming from midbrain tissue meaning the signal from SBE2 may

not be detected. DNA FISH would allow me to perform single-cell analysis on forebrain regions where SBE2 is regulated *Shh* expression.

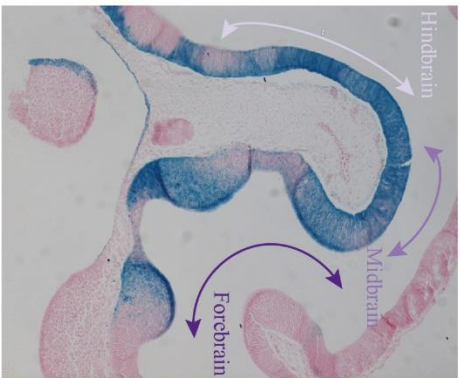
I hypothesised that in developing forebrain tissue, the *Shh* TAD structure will alter in order to bring the distal forebrain enhancers into close proximity with the *Shh* promoter. Chromatin looping may be important for long-range regulation of *Shh* in the forebrain similar to what has been previously observed in the limb bud.

3.3.1 Chromatin conformation of SBE2-regulated forebrain tissue

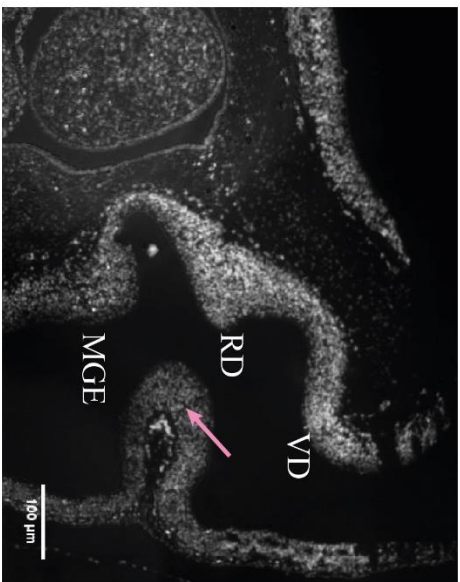
Shh is expressed in the hindbrain, through the midbrain and into the forebrain under the control of several different enhancers (Figure 3.12A). In the forebrain, ventral diencephalic (VD) *Shh* expression is regulated by SBE2 and SBE4; medial ganglionic eminence (MGE) expression is controlled by SBE3 and SBE4; and *Shh* expression in the rostral diencephalon (RD) is regulated by SBE2 and SBE7 (Figure 3.12B). I took sagittal sections of embryo and performed immunofluorescence (IF) against the non-secreted SHH protein on alternate sections to identify RD regions expressing *Shh* (Figure 3.12C). SBE2-driven SHH expression can be observed as a stripe across the RD (Figure 3.12C). Tissue landmarks such as the anterior pituitary gland and the MGE were used in combination with the IF imaging to identify the SBE2 active region in all subsequent experiments. Cells from a forebrain region that never expresses *Shh* are used as a control in all experiments (Figure 3.12B – pink arrow).

Following identification of the SBE2-regulated forebrain tissue, I performed DNA FISH on sections from E10.5 and E11.5 embryos using fosmid probes targeting *Shh*, SBE2 and ZRS (Figure 3.12D). I measured the interprobe distance between all fosmids in both SBE2-regulated and control forebrain cells at both stages of development (Figure 3.13A-D). I found that *Shh* and SBE2 are significantly closer together in SBE2-regulated tissue compared to control (Figure 3.13E). This is consistent at both stages of development although *Shh* and SBE2 are found at even closer proximity at E10.5 (Figure 3.13E). The *Shh*/SBE2 mean interprobe distance of 400-500 nm in control tissue is comparable to what we have previously observed in other tissues e.g. limb bud and mESCs. SBE2 and ZRS are also closer together in SBE2-regulated tissue in comparison to control tissue at both stages of development (Figure 3.13F). This

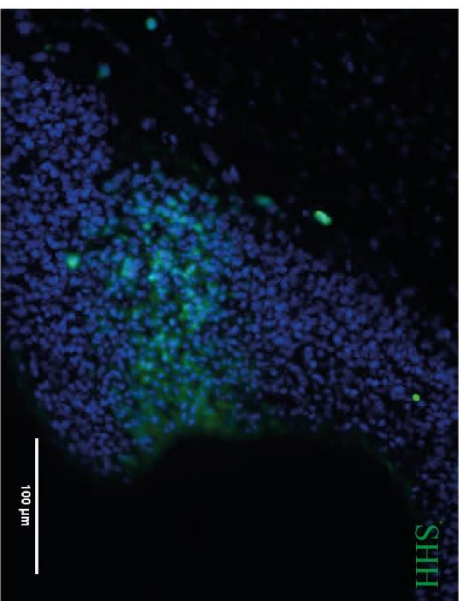
A



B



C



D

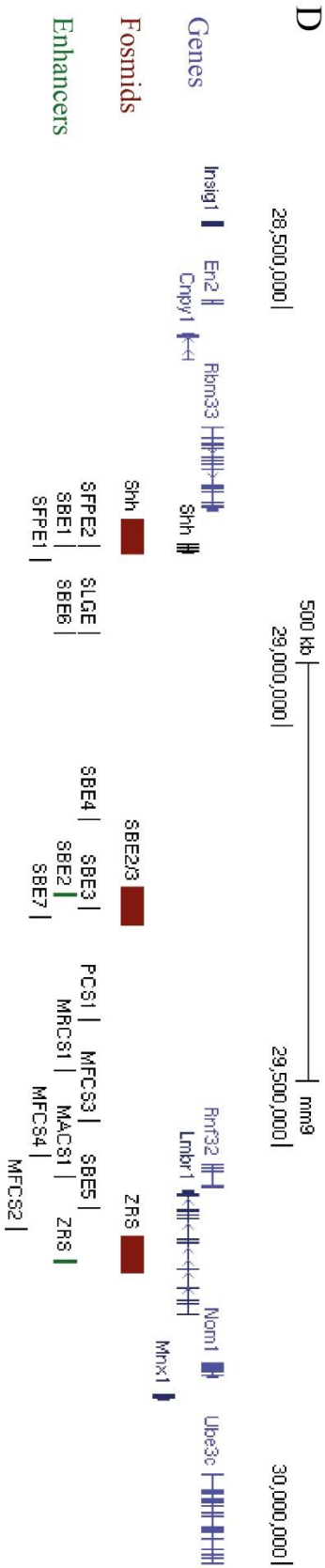


Figure 3.12 Identification of the SBE2-regulated region of the forebrain.

(A) Sagittal head section of an E11.5 embryo stained with β -gal to show expression of a LacZ reporter gene for *Shh*. Arrows indicate the subdivisions of the brain; forebrain, midbrain and hindbrain. (B) Representative image of an E10.5 sagittal section of DAPI stained wild type head. The forebrain consists of the ventral diencephalon (VD), rostral diencephalon (RD) and medial ganglionic eminence (MGE). Control forebrain tissue is indicated (pink arrow). (C) Immunofluorescence showing forebrain regions where *Shh* expression is under the control of SBE2 (Blue = DAPI and green = SHH). (D) ENCODE tracks showing the position of genes (purple), fosmids (red) and enhancers (black, enhancers of interest = green) surrounding the *Shh* locus.

change in conformation may be a consequence of SBE2 coming closer to *Shh* leading to all three loci clustering together (discussed further below). *Shh* to ZRS at E10.5 in SBE2-regulated tissue are significantly further apart (~300 nm) compared to all other tissues (Figure 3.13G). However, one of the biological replicates differs from the other two and seems to be skewing the distribution of interprobe distances. When this is excluded from the analysis there is very little difference in the *Shh*/ZRS interprobe distance in SBE2-regulated and control tissue at both stages of development. *Shh* and ZRS are found close together (~250 nm) irrespective of SBE2 activity and at a distance similar to what we have previously observed in anterior limb tissue and mESCs. Between 10-15% of alleles show co-localisation (<200 nm) between *Shh* and SBE2 and this is specific to SBE2-regulated tissue at both developmental stages (Figure 3.13H). Very low levels of co-localisation are detected in control tissue. This shows that the interactions between these two loci changes specifically in SBE2-regulated tissue and this may be important for *Shh* expression. The frequency distribution of *Shh*-ZRS interprobe distances is unchanged between SBE2-regulated and control tissue (Figure 3.13J). Some changes in the frequency of interprobe distances are observed for SBE2 to ZRS but, again, this is probably an indirect consequence of changes to *Shh* and SBE2 interactions.

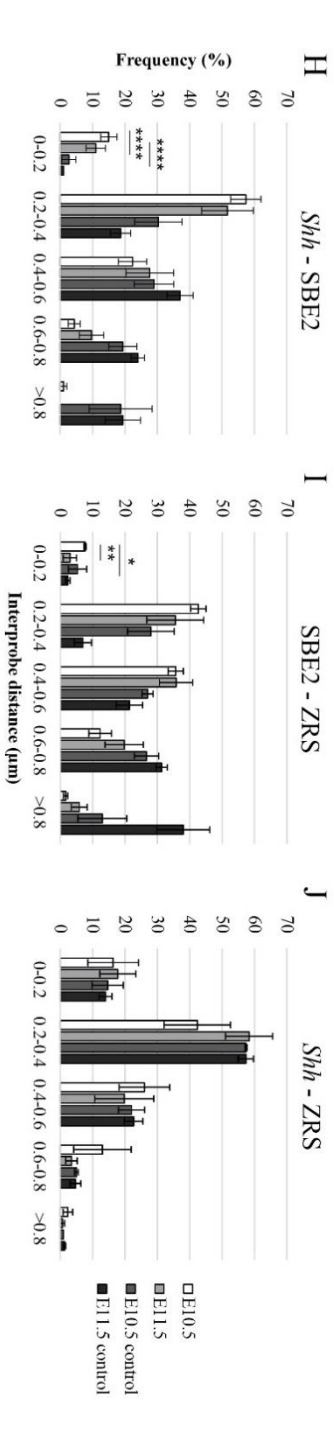
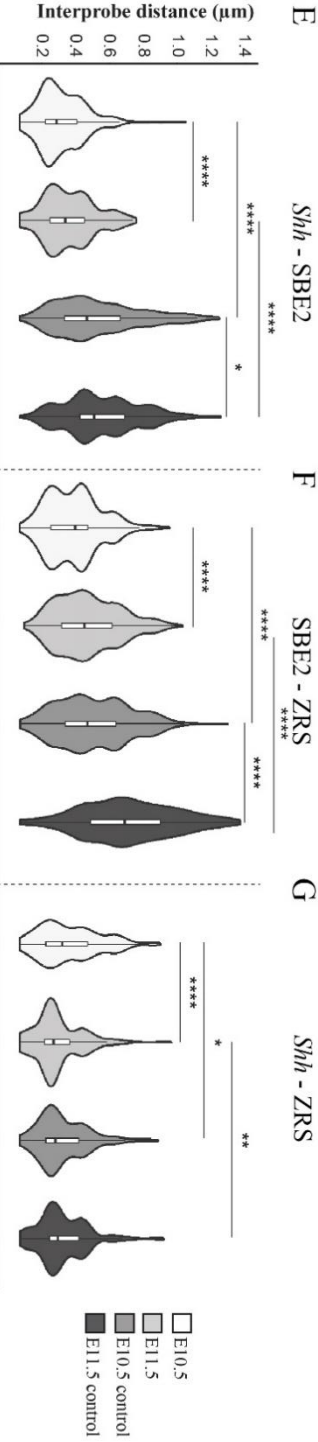
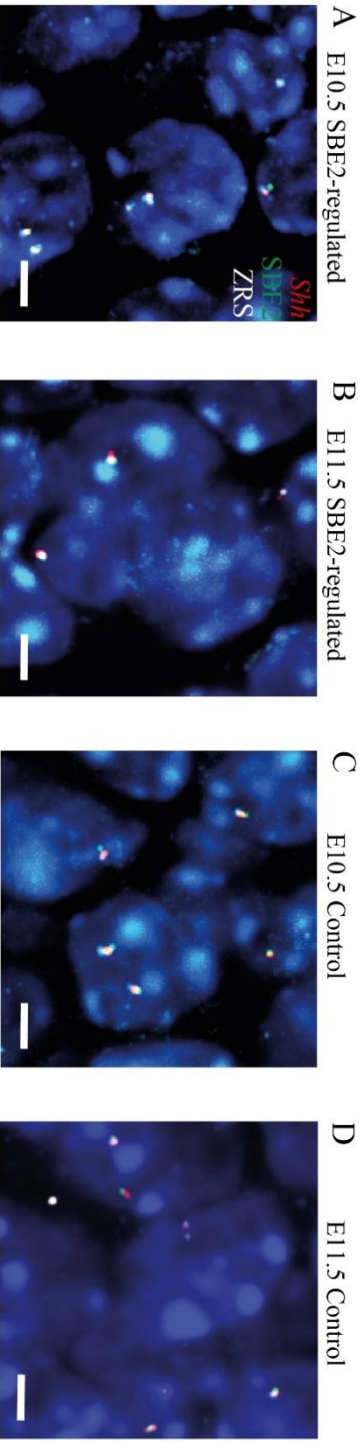


Figure 3.13 *Shh* and its forebrain enhancer show increased proximity in SBE2 active regions of the forebrain.

(A-D) Images of representative nuclei from SBE2-regulated forebrain cells from an (A) E10.5 and (B) E11.5 embryo and from control forebrain cells from an (C) E10.5 and (D) E11.5 embryo showing FISH signals for *Shh*/SBE2/ZRS probes. Scale bars = 5 μ m. (E-G) Violin plots showing the distribution of interprobe distances (μ m) between (E) *Shh*/SBE2, (F) SBE2/ZRS and (G) *Shh*/ZRS probes from forebrain tissue. Boxes show the median and interquartile range of data; n $\approx$ 300. The statistical significance between data sets was examined by Mann-Whitney U Tests, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001. (H-J) Frequency distribution of the interprobe distance calculated every 0.2 μ m bin between (H) *Shh*/SBE2, (I) SBE2/ZRS and (J) *Shh*/ZRS probes comparing SBE2-regulated and control forebrain tissue from E10.5 and E11.5 embryos. Error bars represent SEM obtained from three biological replicates. The statistical significance between data sets was examined by Fisher's Exact Tests, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001.

4-colour DNA FISH enables visualisation of all three fosmid probes targeting *Shh*, SBE2 and ZRS together per chromosome such that the position of each loci relative to other two can be determined simultaneously. *Shh* and ZRS are normally found in close proximity in mESCs and limb tissue. It is unclear in SBE2-regulated tissue where the three loci are relative to each other. In SBE2-regulated tissue of both developmental stages the interprobe distances measured are up to $\sim$ 800 nm and are clustered in lower left hand quadrant (Figure 3.14A & C). Control tissue, in comparison, has an interprobe distance ranging up to $\sim$ 1000-1200 nm and the measurements are far more dispersed (Figure 3.14B & D). Strikingly, there is a clear shift in decreased distance of *Shh* to SBE2 in SBE2-regulated tissue only. This suggests that in SBE2-regulated tissue, the *Shh* TAD adopts a more compact chromatin conformation and this structure may be important for enhancer-promoter interactions.

To look in more depth at the relative arrangement of these three loci, I categorised different possible chromatin arrangements using the interprobe distances (Figure 3.14E). There were very few instances where all three loci are co-localised (<200 nm) or close together (<300 nm) in any of the tissues analysed (Figure 3.14E). When *Shh* and SBE2 are co-localised or adjacent in SBE2-regulated tissue, ZRS is >200 nm away at both stages of development. There are significantly more conformations where ZRS is within 300 nm of either *Shh* or SBE2 in control tissue (Figure 3.14E) in line with

previous data where *Shh* and ZRS are found close together in non-expressing limb tissue. These data suggest that in SBE2-regulated tissue the organisation of the *Shh* TAD alters allowing an interaction between *Shh* and SBE2 to be the favoured. *Shh* and ZRS do not need to be in close proximity for SBE2 regulation. All three loci are rarely found <300 nm apart in SBE2-regulated tissue (~5%) and it seems when *Shh* is close to SBE2 the ZRS is further away. In comparison, the default conformation of *Shh* and ZRS being in close proximity is observed in control tissue.

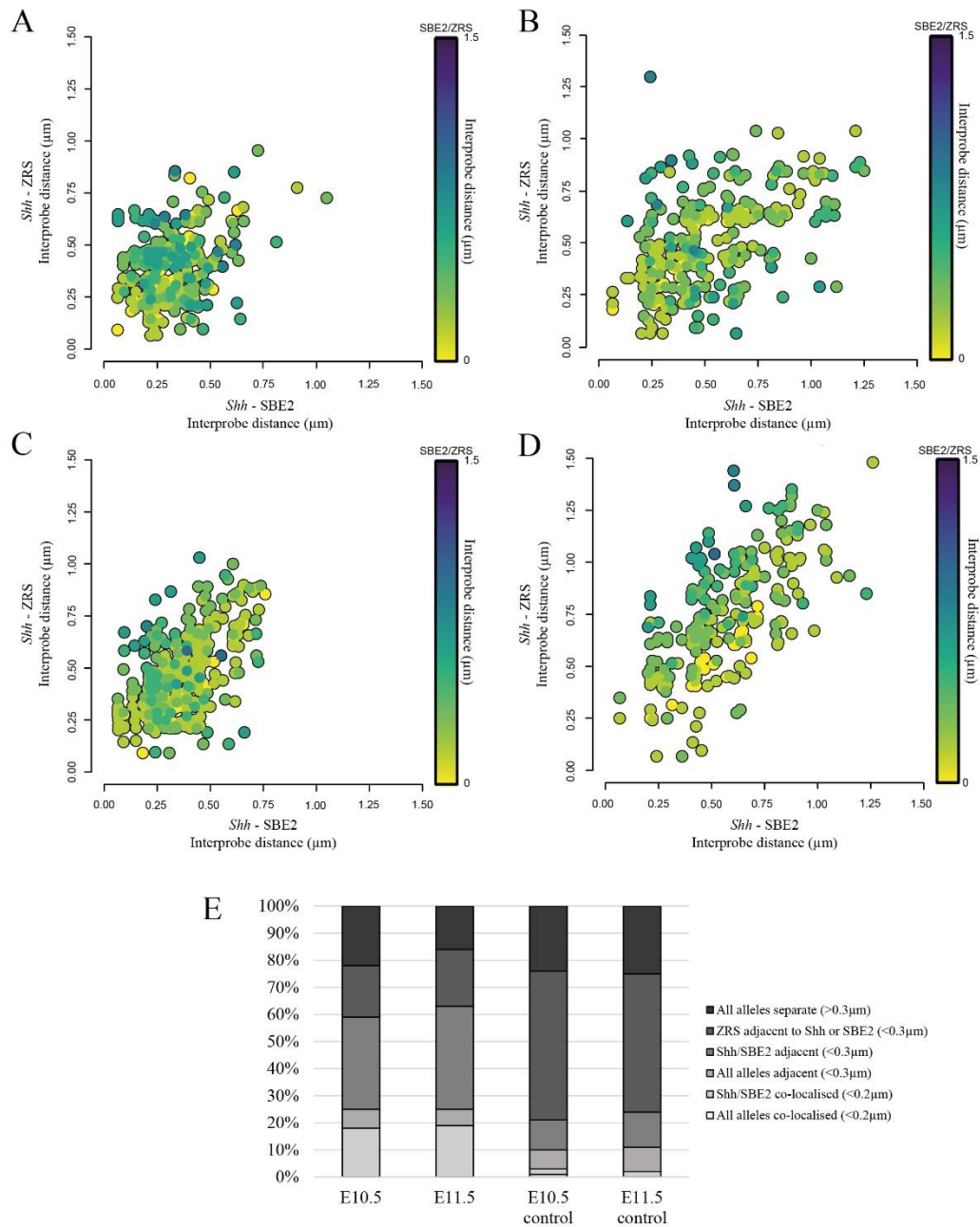


Figure 3.14 The *Shh* TAD is more compact in *Shh*-expressing tissue.

(A-D) Scatter plots showing the distribution of distances between *Shh*/SBE2 (x-axis), *Shh*/ZRS (y-axis) and SBE2/ZRS (colour chart) from SBE2-regulated forebrain cells from an (A) E10.5 and (B) E11.5 embryo and control forebrain cells from an (C) E10.5 and (D) E11.5 embryo. Each point represents all three distances measured at a single allele; n~300. (E) Stacked bar chart showing the proportion of different chromatin conformations between *Shh*/SBE2/ZRS in forebrain tissue. All data obtained from three biological replicates.

3.3.2 Development of an RNA FISH assay to measure the proportion of expressing alleles in embryonic tissue

We have previously used methods such as whole-mount *in situ* hybridisation (WISH) and qRT-PCR to visualise and measure levels of *Shh* expression in the developing limb bud (Lettice et al., 2014; Williamson et al., 2019). WISH is a useful method for visualising *Shh* spatial expression patterns particularly for looking at disruption to expression patterns in mutant mouse models, but it cannot be used quantitatively. *Shh* quantification by qRT-PCR is difficult to do on dissected embryonic tissue as samples can be contaminated by non-expressing cells. This process is even more difficult to do on complex tissue such as the forebrain which is embedded inside the embryo. Therefore, I aimed to develop an RNA FISH assay to visualise and quantify *Shh* expression at the single-cell level in complex embryonic tissue. I designed RNA FISH probes targeting the nascent transcript of *Shh* using Stellaris® RNA FISH Probe Designer (Figure 3.15A). I targeted the nascent transcript as I wanted to ultimately use the RNA probes in combination with DNA fosmid to look at chromatin structure at a *Shh*-expressing chromosome. The RNA FISH spots will appear as either 1 or 2 signals per cell indicating *Shh* expression from either 1 or 2 alleles. I performed RNA FISH on a wild type E11.5 embryo and imaged the neural tube and notochord as these were two areas of tissue that express high levels of *Shh* throughout the entire embryo.

Shh is expressed in the floorplate and is important for dorso-ventral patterning of the neural tube. *Shh* expression can be detected in almost every cell of the floorplate (Figure 3.15B), whereas at the dorsal side of the neural tube of the same tissue section no *Shh* signal is detected (Figure 3.15C). I quantified *Shh* expression in the floorplate of E10.5 embryos and found *Shh* was expressed from 49% of alleles (Benabdallah et al., 2019). Notochord cells also show *Shh* signal from the majority of cells (Figure 3.15D). No *Shh* expression is detected in any of the cells surrounding the notochord. The tissue section also included some gut epithelial linings where *Shh* is expressed. *Shh* signal can also be detected in these cells albeit at lower levels than the notochord or floorplate cells (Figure 3.15E). *Shh* signal is never detected in cells surrounding any of the tissues of interest. RNA FISH can be used to robustly detect varying levels of *Shh* expression across different tissues of the developing embryo.

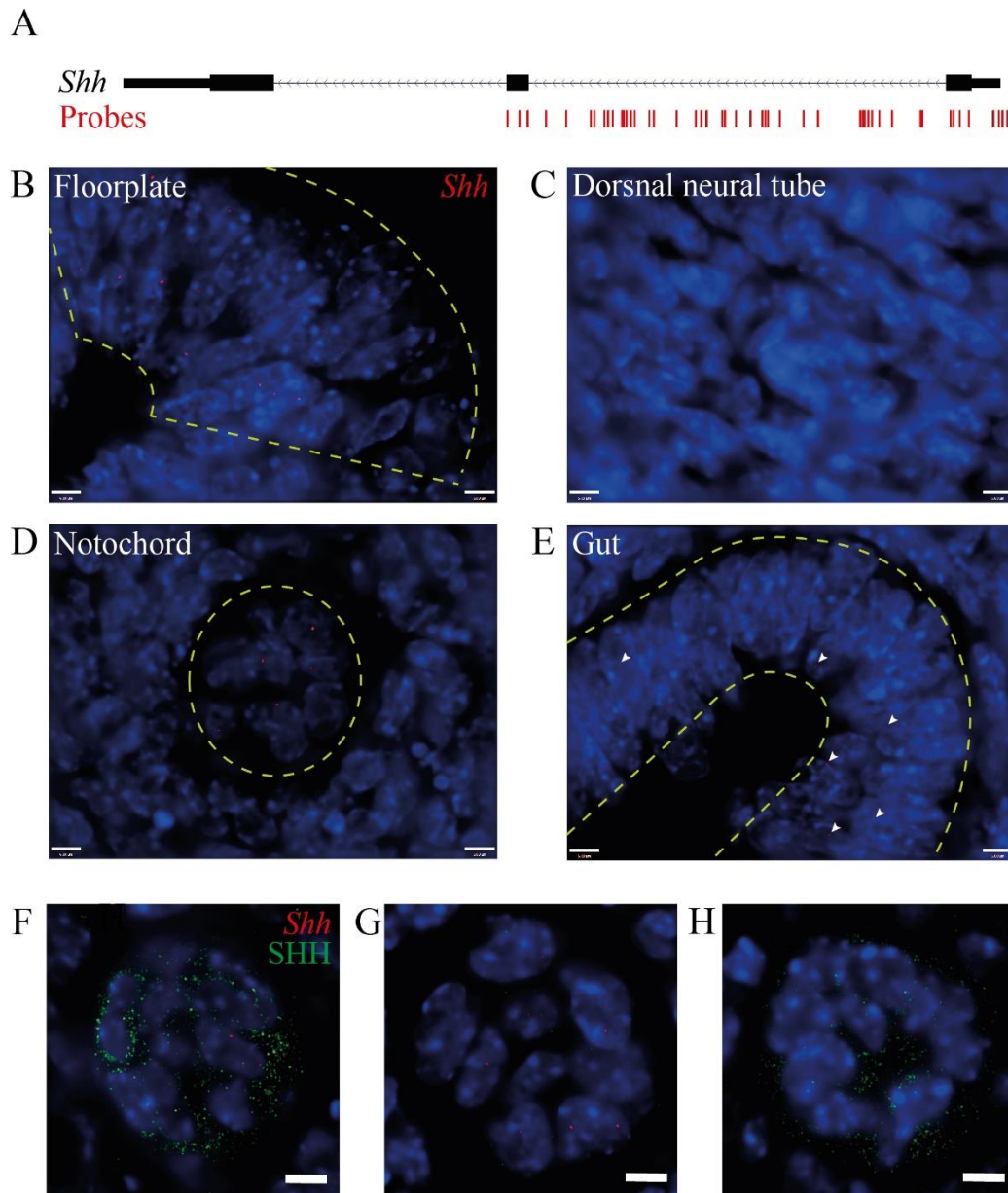


Figure 3.15 Visualising *Shh* nascent expression using oligo probes.

(A) Schematic of the *Shh* gene showing the position of the directly labelled oligo probes (red). (B-E) Representative images of nuclei taken from the (B) floorplate, (C) dorsal neural tube, (D) notochord and (E) gut epithelial lining of an E11.5 embryo showing RNA FISH signals for the nascent *Shh* transcript. The tissue of interest is contained within yellow dashed boundary. White arrowheads indicate RNA FISH signal. Scale bars = 6 µm. (F-H) Images of representative nuclei from the notochord of an E11.5 embryo following immuno-RNA FISH targeting SHH (green) and *Shh* nascent transcript (red). Images were taken following (F) the full protocol, (G) omitting the antigen retrieval step or (H) treating slides with RNase.

I combined the RNA FISH assay with immunofluorescence to confirm that the RNA signal is only detected in cells that express *Shh*. Using an antibody that detects the non-secreted c-terminal half of the SHH protein, I found that in notochord *Shh* RNA is detected in the nucleus of cells while SHH protein is found in the cytoplasm of these cells (Figure 3.15F). When the antigen retrieval step is omitted from the immuno-RNA FISH protocol, no SHH protein is detected whilst the RNA signal persists (Figure 3.15G). Slides treated with RNase before probe hybridisation have no *Shh* RNA signal but SHH can still be detected in the cytoplasm of cells (Figure 3.15H). From these experiments, I concluded that the RNA FISH probes can be used to specifically detect *Shh* expression in tissue sections at the single cell level.

3.3.3 Expression analysis of SBE2 regulated forebrain tissue

I proceeded to quantify the number of expressing alleles in the SBE2-regulated region of the forebrain at both E10.5 and E11.5 stages by RNA FISH (Figure 3.16A & B). Approximately 60% of alleles are expressing *Shh* at both stages of development (Figure 3.16C). This is slightly higher expression levels than what I have previously measured in floorplate cells (Benabdallah et al., 2019). I also looked at the proportion of cells expressing *Shh* from one or two alleles (Figure 3.16D). Approximately 40% of cells are expressing *Shh* from one allele and 20% are expressing from both alleles (Figure 3.16D). There is no difference in expression detected between the embryonic stages.

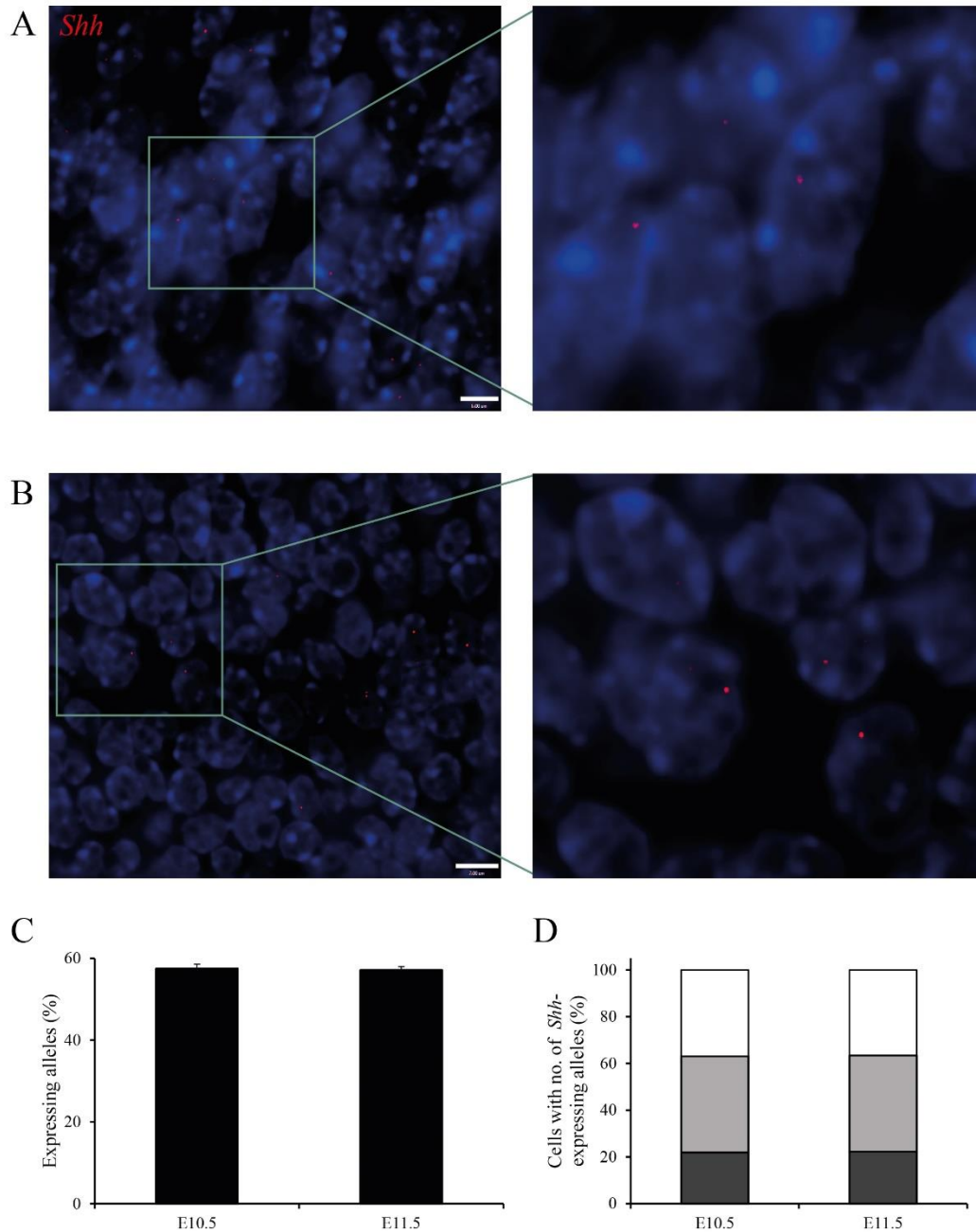


Figure 3.16 *Shh* is expressed at similar levels in the forebrain across developmental stages.

(**A & B**) Representative images of nuclei taken from the SBE2-regulated region of the forebrain of an (**A**) E10.5 and (**B**) E11.5 embryo showing RNA FISH signals for *Shh* nascent transcript (red). A zoomed in image of some nuclei are shown on the right hand panels. (**C**) Quantification of the percentage of *Shh* expressing alleles in SBE2-regulated forebrain tissue assayed by RNA FISH. (**D**) Proportion of SBE2-regulated forebrain cells expressing *Shh* from 1, 2 or neither chromosome.

3.4 Discussion & conclusions

In this chapter, I used ATAC-seq to study chromatin accessibility in *Shh*-expressing cells isolated from dissected fore-/midbrain across different stages of embryonic development. I showed that the accessibility of the *Shh* regulatory landscape is relatively unchanged in *Shh*-expressing versus non-expressing cells and across development. I then used motif analysis to identify potential transcription factors that are restricted to *Shh*-expressing cell populations at particular stages of development that may function to regulate patterning of the embryonic fore-/midbrain and neuronal differentiation. I next focused on the forebrain enhancer SBE2 and used DNA FISH to demonstrate that this enhancer is found in closer proximity to the *Shh* promoter specifically in SBE2-regulated tissue. Finally, I developed a nascent RNA FISH assay that was be used to quantify *Shh* expression in embryonic tissue at the single-cell level.

3.4.1 Use of a *Shh* reporter mouse model to study enhancer regulation

Previous work in the lab has used dissected embryonic limb tissue to study which transcription factors bind the ZRS limb enhancer, what interactions occur across the *Shh* TAD, and which chromatin modifications are present in different regions of the limb (Lettice et al., 2012; Peluso et al., 2017; Williamson et al., 2016, 2019). Techniques such as ChIP and 5C have provided useful insights into the function and regulation of the ZRS enhancer but require relatively large quantities of cells (millions). *Shh*-regulated brain tissue is embedded within the head of the embryo and virtually impossible to dissect out any region of interest without contaminating the sample with cells from the surrounding regions. Here instead, I used a *Shh* reporter mouse model which expressed GFP under the control of the *Shh* promoter to sort *Shh*-expressing cells from non-expressing cells. I could also dissect out fore-/midbrain tissue from the embryo more consistently between samples as I was targeting a larger area of the embryo. I only obtained between 8,000-30,000 *Shh*-expressing cells per FAC-sorted embryo meaning that many embryos would need to be pooled in order to perform techniques such as ChIP or 5C. I decided to focus on chromatin accessibility in *Shh*-expressing cells derived from the fore-/midbrain tissue of the reporter mouse model.

3.4.2 Chromatin accessibility marks enhancers but not their activity

I was surprised to find that *Shh* enhancers SBE1, SBE5 and SBE7 were accessible in both *Shh*-expressing and non-expressing cells. In the limb bud, the ZRS enhancer is poised but inactive across the entire distal tip and additional H3K27 acetylation is needed to specifically activate the ZRS in the ZPA (Peluso et al., 2017). A similar model could apply to the *Shh* neural enhancers. These enhancers may also be open, poised but inactive across all cells of the embryonic head earlier during development and later become specifically activated in a spatially-restricted manner. It would be useful to apply methods that require low cell numbers such as CUT&RUN/CUT&TAG (Kaya-Okur et al., 2019; Skene et al., 2018) to look at the histone marks H3K4me1 and H3K27ac in the different cell populations isolated from the *Shh* reporter mouse model. This would provide important insights into the chromatin dynamics of these fore-/midbrain enhancers. It would also be interesting to see whether the other forebrain enhancers that are not found to be accessible (SBE3/SBE4) or have low levels of accessibility (SBE2) are marked by these chromatin modifications.

I also looked at the chromatin accessibility across the regulatory landscape of two other developmental genes; *Otx2* and *Gbx2*. *Otx2* is expressed in the epiblast, anterior neuroectoderm, forebrain and midbrain under the control of enhancers with distinct temporal activities (Kurokawa et al., 2004a, 2004b, 2014). The AN1 enhancer is active at E7.5-8.5 whereas FM1 is active between E9.5-10.5 (Kurokawa et al., 2004a, 2004b). I found that the AN1 enhancer was accessible across all stages of development analysed despite the activity of this enhancer being decommissioned at E8.5. Why some enhancers remained accessible and when they are ‘closed’ is unclear. Perhaps some enhancers are used again later in development or postnatally for other functions. FM2 enhancer (120 kb downstream of *Otx2*) is also active in fore-/midbrain tissue between E9.5-10.5 and no peak of accessibility is observed in this region or for any of the downstream retinal enhancers (Chan et al., 2020; Kurokawa et al., 2004b). Similarly to the *Shh* enhancers SBE3 and SBE4, FM2 shows *in vivo* enhancer reporter activity. One potential explanation could be that these enhancers function in a very small cell population and the ATAC-seq signal may not be sufficiently enriched above background.

Gbx2 is a transcription factor spatially restricted to the anterior end of the hindbrain and functions to specify the MHB (Wassarman et al., 1997). I found that *Gbx2* and its potential enhancer (Islam et al., 2006) are accessible in all ATAC-seq cell populations despite these cells being derived from fore-/midbrain tissue. There is a possibility that some hindbrain cells may have contaminated the samples during the dissection process. However, together with the results from the *Shh* and *Otx2* regulatory landscapes I believe that at least a subset of enhancers become accessible earlier during development prior to their activation. Upon activation, these enhancers can then direct specific spatio-temporal expression patterns of their target gene. These enhancers remain accessible even when no longer active. These data suggest that chromatin accessibility can be a useful indication of the presence of an enhancer sequence but does not provide information about its spatial or temporal activities. More information regarding other chromatin modifications are required in order to determine this.

3.4.3 Identification of temporally-regulated transcriptional networks in *Shh*-expressing cells from neural tissue

The motifs predicted to be bound by RFX or SOX transcription factors showed an interesting spatio-temporal pattern. RFX/SOX motifs were enriched early in E10.5 *Shh*-expressing cells before subsequently being lost in *Shh*-expressing cells at E11.5. Instead non-expressing cells at E11.5 became enriched for RFX/SOX motifs. Enrichment for RFX/SOX motifs were later gained in *Shh*-expressing cells only at E14.5. It would be of interest to follow up whether this dynamic pattern of RFX/SOX regulation can be related to any members of these transcription factor families. This could be achieved by performing *in situ* hybridization or immunofluorescence experiments across these stages of embryonic development to follow the expression pattern of different *Rfx* or *Sox* genes. However, low reproducibility was found between replicates for E11.5 *Shh*-expressing samples and this analysis should be repeated using only common reproducible peaks in the future.

Otx2 is an important regulator of both early and late forebrain and midbrain development and together with *Gbx2* helps to establish the midbrain-hindbrain boundary (MHB) or isthmus organizer, an important signaling center during mid-

/hindbrain development (Arenas et al., 2015). I found that OTX2 binding motifs were absent in accessible chromatin regions of *Shh*-expressing cells at the earlier stages of development. It has been previously shown that *Otx2* expression at E10.5 only overlaps with *Shh* in the midbrain and expression is not detected in *Shh*-expressing forebrain tissue (Simeone et al., 1995). These authors also demonstrated that *Otx2* expression could be repressed by retinoic acid. I propose that at E10.5, retinoic acid binding at *Otx2* represses expression whilst inducing *Shh* expression establishing and maintaining the posterior identity of these cells. This may be important for establishing the anterior-posterior axis of the developing forebrain. In the developing limb bud, retinoic acid signaling is also important for polarization and induces *Shh* from an exogenous source (Tickle and Towers, 2017).

Another temporally-regulated transcription factor found exclusive to *Shh*-expressing cells was TCF3. *Tcf3* is a regulator of neuronal differentiation and is important for maintaining neural stem cell identity. The differentiation of neural stem cells to neural progenitors is regulated by WNT signaling (Inestrosa and Varela-Nallar, 2015). Using GO analysis, I found that *Shh*-expressing cell populations from E11.5 were involved in processes such as neuronal differentiation and WNT signaling regulation. Perhaps at E11.5, when *Tcf3* expression levels peak, WNT signaling is repressed in *Shh*-expressing cells preventing neuronal differentiation. WNT repression may also be important in these cells to establish proper dorsal-ventral patterning.

3.4.4 Assessment of the robustness of ATAC-seq datasets

Many of the replicate ATAC-seq samples were inconsistent relative to other replicates as exemplified by the low IDR scores (range 8.8-65.4) suggesting there may be some technical issues with sample acquisition. There will be some variability in samples due to differences between embryos, litter and developmental stage when dissecting forebrain/midbrain tissue meaning that some samples could potentially be contaminated with hindbrain tissue to varying degrees. The lack of a pure *Shh*-expressing cell population only from forebrain/midbrain tissue could explain why there is large variability between biological replicates. The data quality of the majority of the ATAC-seq dataset is low meaning that the data must be interpreted with caution

and no conclusions should be drawn until a more robust dataset is generated. I was unable to observe ATAC peaks corresponding to known *Shh* forebrain enhancers (e.g. SBE2) which again, may be due to the lack of a pure cell population where this enhancer is active. One method to achieve a more pure *Shh*-expressing cell population from forebrain and/or midbrain tissue would be to include an additional marker e.g. second fluorescent signal, in the mouse model that would mark forebrain and/or midbrain cells specifically allowing these cells to be isolated by FACS without the need for crude dissections.

3.4.5 Distal forebrain enhancer of *Shh* is found in close proximity

Others in the group had shown previously that the ZRS limb enhancer shows increased co-localisation with the *Shh* promoter in ZRS-regulated tissue (Williamson et al., 2016), whereas the *Shh* neural enhancer SBE6 shows increased separation from the *Shh* promoter upon activation (Benabdallah et al., 2019). This suggests that these enhancer may use different modes of enhancer-promoter communication e.g. chromatin looping vs phase condensates (1.2.2). I found that the *Shh* forebrain enhancer SBE2 also shows increased proximity and co-localisation to the *Shh* promoter in SBE2-regulated tissue. This suggests that SBE2 may depend on a chromatin looping mechanism similar to ZRS. The ZRS is flanked by CTCF sites that function to direct this chromatin looping interaction (discussed in subsequent results chapters) whereas no bound CTCF sites are found near SBE2. It is unclear what drives the interaction between *Shh* and SBE2 in forebrain tissue.

Shh and ZRS are found in close proximity in both *Shh*-expressing (ZPA) and non-expressing (anterior) limb bud tissue. This preformed topology between *Shh*/ZRS has been suggested to help facilitate ZRS-mediated *Shh* expression in the limb bud (Paliou et al., 2019; Williamson et al., 2016). It would be interesting to look at the chromatin interactions between *Shh* and SBE2 at earlier stages of development (E8-9.5) prior to SBE2 activation to see if a pre-formed interaction also exists between these loci or does the interaction coincide with enhancer activation.

One of the important questions in the field is whether enhancer-promoter close proximity coincides with transcription. To address this question I intended to develop an RNA-DNA FISH assay to test whether *Shh*/SBE2 proximity is coupled to nascent *Shh* transcription. The RNA FISH assay for *Shh* nascent transcription works well, however, preliminary experiments trying to combine this with DNA FISH have thus far been unsuccessful. The RNA signal appears to be destroyed at some point during the DNA FISH protocol and further work is needed. RNA-DNA FISH for the *Shh* locus on tissue sections have been previously carried out on limb tissue (Amano et al., 2009) thus, in the future myself or others in the group may be able to test it on SBE2-regulated forebrain tissue.

3.4.6 SBE7-mediated gene regulation in the developing forebrain

The fosmid used in the DNA FISH experiment to target the SBE2 forebrain enhancer also binds to a region that encompasses the SBE3 and SBE7 enhancers. SBE3 has been shown to be active in the MGE region of the forebrain and should not be active in the tissue used here for the DNA FISH experiments and is unlikely to have any effect on *Shh* TAD structure (Jeong et al., 2006). In contrast, SBE7 has recently been identified as a regulator of *Shh* expression in the precordal plate and diencephalic region of the forebrain that overlaps with SBE2 activity (Crane-Smith et al., 2019; Sagai et al., 2019). I therefore cannot rule out the increased interactions observed in the forebrain tissue may be alternatively driven by either SBE7 alone or in combination with SBE2. In the future, the use of oligo probes for DNA FISH would provide an opportunity to distinguish between both enhancers and look at the chromatin interactions within tissue.

Chapter 4. The role of CTCF in *Shh* distal regulation

4.1 Introduction

CTCF is one of the main architectural proteins implicated in the structuring of the genome into TADs (Narendra et al., 2015; Rao et al., 2014; Sanborn et al., 2015). Convergent orientated CTCF binding sites are enriched at TAD boundaries and at the base of the chromatin loops in mammalian genomes (Rao et al., 2014). CTCF forms chromatin loops and TADs through a cohesin-driven loop extrusion process (1.1.2). CTCF was also demonstrated to act as an insulator by blocking ectopic enhancer interactions and its' function as an insulator and architectural protein has linked it to gene regulation.

TADs are proposed to reduce the search space of enhancers for their target promoter whilst CTCF binding at the boundaries is thought to insulate these enhancers from other genes (Fudenberg et al., 2016; Sun et al., 2019). The formation of chromatin loops can bring enhancers into close proximity to their target gene and this looping is one of the main models of enhancer-promoter communication (1.2.2.2). When constrained regulatory interactions are disrupted by mutations that disrupt TAD boundaries this can have severe pathological consequences (Flavahan et al., 2016; Franke et al., 2016; Lupiáñez et al., 2015). A number of studies have demonstrated for individual loci that structural variants can disrupt TADs and re-wire enhancer-promoter interactions, linking TAD structure and gene regulation (Flavahan et al., 2016; Franke et al., 2016; Lupiáñez et al., 2015). However, these studies involve removing large regions of DNA with the disruption of multiple DNA elements and not just CTCF binding sites. Genome-wide depletion of proteins involved in loop extrusion (CTCF, cohesin complex, NIPBL) leads to the disruption of TADs whilst having limited effects on gene expression (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). These conflicting results have brought into question the importance of TADs as units of gene regulation.

Our lab groups have previously identified a number of prominent CTCF binding sites at the *Shh* TAD boundaries in the convergent orientation (Figure 4.1) (Williamson et al., 2019). To test if these CTCF binding regions are important for the formation of the *Shh* TAD and maintaining the enhancer-promoter spatial proximity observed across

various cell/tissue types, CRISPR-Cas9 genome editing was used (Laura Lettice, Hill group) to delete each of these major CTCF peaks (~1 kb) in mouse E14TG2A embryonic stem cells (mESCs) (Williamson et al., 2019). These gRNAs were also used to generate homozygous Δ CTCF mutant mice (carried out by Laura Lettice). *Shh* TAD structure, regulatory interactions and developmental gene expression patterns were then assessed in Δ CTCF mutant mESCs and embryonic mouse tissue. The main findings of this chapter are included in our publication (Williamson et al., 2019).

4.1.1 CTCF binding sites deletion alters *Shh* TAD structure in mESCs

In this section, the 5C and DNA FISH experiments in mESCs discussed were carried out by Iain Williamson (Bickmore Group) and I assisted with the DNA FISH analysis. For more details please refer to Williamson et al., (2019) which is included in 8.2.

4.1.1.1 *Shh* TAD boundary

As discussed previously, the *Shh* gene is flanked by two CTCF sites (denoted as CTCF1 at the 3' end and CTCF2 at the 5' end of the gene) marking the *Shh* TAD boundary (Figure 4.1). To characterise the *Shh* TAD structure and chromatin conformation following removal of either of these CTCF sites, 5C and DNA FISH was performed on homozygous Δ CTCF1 and Δ CTCF2 mESCs lines (Williamson et al., 2019). 5C is a high-throughput version of 3C and allows the interrogation all chromatin interactions throughout the *Shh* TAD including the neighbouring TAD regions. *Shh* expression is suppressed by Polycomb in mESCs so this allowing the assessment of the TAD structural changes following CTCF site knockout in the absence of expression.

In Δ CTCF1 cells, the 5C analysis conducted by Iain Williamson shows that *Shh* has lost interactions with the rest of its' own TAD, and the TAD boundary has shifted proximal to the CTCF2 site (Williamson et al., 2019). There is also increased inter-TAD interactions with the neighbouring *En2* TAD and loss of intra-TAD interactions with the *Shh* sub-TAD boundary (Williamson et al., 2019). Δ CTCF2 cells have an enriched boundary at CTCF1 that appears to compensate for loss of CTCF2 TAD-wide interactions and, similarly to Δ CTCF1 cells, shows increased inter-TAD interactions with the *En2* TAD. Intra-TAD interactions become stronger across the *Shh* sub-TAD

(Williamson et al., 2019). This suggests that CTCF2 normally participates in forming the sub-TAD loop and loss of this structure means the sub-TAD is now unstructured and sub-TAD regions are interacting more randomly. The TAD boundary at *Lmbr1* is unaffected in both these knockout cell lines (Williamson et al., 2019). CTCF1 normally marks the TAD boundary but this data shows that both CTCF sites 1 and 2 can function redundantly as boundaries and participate to block cohesin driven loop extrusion. Both knockout cell lines showed altered chromatin conformations with a significant increase in inter-probe distances across the TAD consistent with the 5C data and suggestive of a decompaction or re-structuring of the *Shh* TAD (Williamson et al., 2019). In Δ CTCF1 cells, a decreased distance between *Shh* and *Cnpy1*, located in the *En2* TAD, was found consistent increased inter-TAD interactions observed by 5C (Williamson et al., 2019).

4.1.1.2 *Lmbr1* TAD boundary

At the other end of the *Shh* TAD there are three prominent CTCF sites (Figure 4.1). CTCF3 is found 3' of the ZRS limb enhancer within intron 5 of *Lmbr1* and this site interacts strongly with CTCF1/2 at the *Shh* boundary forming a chromatin loop in all cells and tissues analysed thus far (Paliou et al., 2019; Williamson et al., 2019). CTCF4 and 5 are at the 5' end of *Lmbr1* promoter with CTCF4 found the most proximal at the TAD boundary between the *Shh* TAD and the neighbouring *Mnx1* TAD. I will refer to this region encompassing CTCF sites 1-3 as the *Lmbr1* TAD boundary. As in 4.1.1.1, 5C and DNA FISH was performed on CTCF knockout mESCs to characterise the *Shh* TAD structure (Williamson et al., 2019).

Loss of the chromatin loop in Δ CTCF3 cells is compensated by a gain of interactions between CTCF1 site and the ZRS (Williamson et al., 2019). It has been shown that deletion of CTCF3 results in ectopic CTCF binding at the ZRS (Paliou et al., 2019). Δ CTCF3 cells showed enriched sub-TAD interactions by 5C similar to those observed for Δ CTCF2 (Williamson et al., 2019). 5C heatmaps from Δ CTCF4 cells are very similar to wild type cells with some increased interactions from CTCF3 detected (Williamson et al., 2019). No change in boundary positions was seen in Δ CTCF3, 4 or 5 cells (Williamson et al., 2019). Δ CTCF5 cells show a weakened *Lmbr1* TAD

boundary that has shifted closer to *Nom1* and CTCF5 loses its interactions with the adjacent *Mnx1* TAD ((Williamson et al., 2019). Similar to the results in Δ CTCF1/2, increased inter-probe distances were observed across the TAD in Δ CTCF3 cells (Williamson et al., 2019). In Δ CTCF4 cells, only an increased distance from SBE2 to ZRS and in Δ CTCF5 only from *Shh* to ZRS (Williamson et al., 2019). ZRS is found closer to *Mnx1* in Δ CTCF4 cells by DNA FISH (Williamson et al., 2019).

Overall, changes to intra-TAD interactions, TAD boundary positions and inter-probe distances across the *Shh* TAD are seen in Δ CTCF mESCs. Loss of CTCF1-3 have a more profound effect on *Shh* TAD structure and chromatin conformation as demonstrated by 5C and DNA FISH suggesting that the peak interaction or chromatin loop observed between *Shh*/ZRS is most likely formed by these CTCF sites. CTCF1 and 5 seem to be more important for insulating the *Shh* TAD from neighbouring TADs and removal of these sites affects inter-TAD interactions.

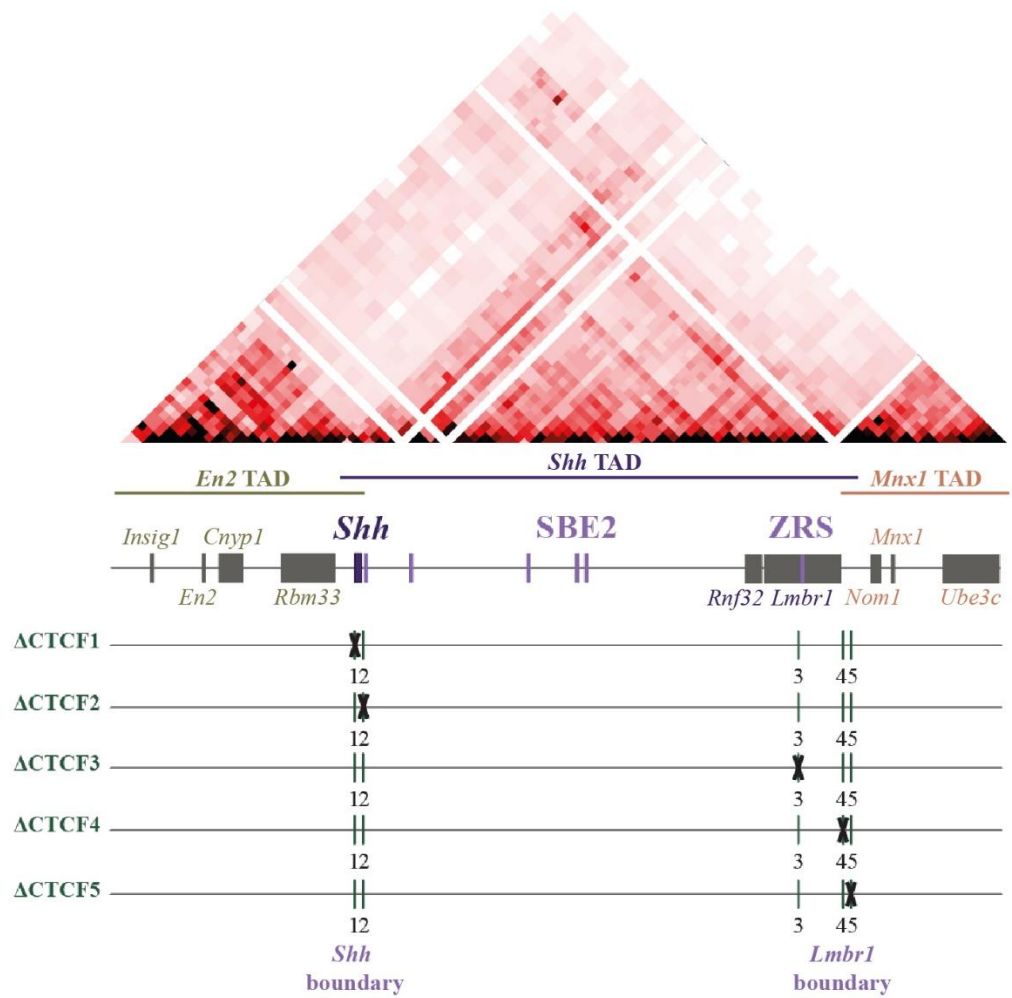


Figure 4.1 Removal of CTCF binding sites at the *Shh* TAD boundaries.

Representative 5C heatmap from a whole E11.5 embryo (provided by Iain Williamson) showing the *En2*, *Shh* and *Mnx1* TADs, the location of genes in each TAD (grey) with the *Shh* gene, enhancers and boundary positions indicated (purple). Below are the positions of CTCF binding sites 1-5 (green) showing the knockouts generated for each site (Δ CTCF).

4.2 Tissue-specific enhancer-promoter proximity is lost in Δ CTCF mutants

In this section, the experiments discussed were carried out by me with assistance from Iain Williamson (Bickmore Group) with the DNA FISH analysis of limb tissue.

4.2.1 *Shh*-SBE2 in forebrain tissue

Both 5C and DNA FISH approaches demonstrated that TAD organisation is disrupted in the Δ CTCF mESCs (Williamson et al., 2019). I was interested in the effect that loss of CTCF binding sites at the *Shh* TAD boundaries and TAD perturbation would have on distal enhancer regulation in a developmental context as CTCF has been implicated in enhancer-promoter looping. To characterise the spatial proximity of *Shh*, SBE2 and ZRS in Δ CTCF mutant tissue, I performed 4-colour 3D FISH on tissue from E11.5 wild type and Δ CTCF embryos (Figure 4.2). Here, I discuss results from measurements taken in SBE2-regulated *Shh*-expressing regions of the forebrain and compared to those from non-expressing control forebrain tissue (3.3.1).

SBE2-regulated regions in the rostral diencephalon (RD) of the forebrain were identified in E11.5 tissue by immunofluorescence staining for SHH (Figure 4.2A & B). Similar to 3.3.1, cells from a forebrain region that never expresses *Shh* were used as a control in all experiments (Figure 4.2A – pink arrow). With the exception of CTCF5, increased spatial proximity between *Shh*/SBE2 is observed in SBE2-regulated forebrain tissue from the Δ CTCF mutants compared to wild type tissue (Figure 4.2D) and the distances between *Shh*/SBE2 in Δ CTCF mutants become similar to those observed in control tissue (compare left hand panels in Figure 4.2D & E). The distances between *Shh*/ZRS are mostly unchanged in Δ CTCF mutants compared to wild type with the exception of CTCF4 where *Shh* and ZRS are found significantly closer together (Figure 4.2D). There are also some changes in the *Shh* to SBE2 interprobe distances in SBE2-regulated tissue from Δ CTCF mutants (Figure 4.2D).

In the control forebrain tissue, more subtle effects are seen in the Δ CTCF mutants with regards to *Shh*/SBE2 and *Shh*/ZRS distances with the predominant effects observed in the distances between SBE2/ZRS loci (Figure 4.2E). Despite the changes in *Shh*/SBE2 spatial proximity seen across all Δ CTCF mutant forebrain tissue the mutant mice are

phenotypically normal (Williamson et al., 2019). This indicates that although there is decreased *Shh*/SBE2 proximity, regulation of *Shh* expression via SBE2 is either not disrupted or is disrupted at a level too subtle to induce phenotypic consequences. Analysis of SBE2 knockout mice (carried out by another PhD student Zoe Crane-Smith) found only a mild phenotype including malformed pituitary glands and patterning defects (Crane-Smith et al., 2019). Potentially SBE2 can still contact the *Shh* promoter but in the absence of CTCF binding at the TAD boundary this interaction is more transient or less frequent and is not apparent above background within this assay. Alternatively, SBE2 may not require direct contact with the *Shh* promoter in order to regulate its expression and could use an alternative mechanism of enhancer-promoter communication.

4.2.2 *Shh*-ZRS in limb tissue

Limb bud expression of *Shh* is solely determined by the ZRS enhancer which co-localises with the *Shh* gene specifically in posterior limb bud cells where this enhancer is active (Lettice et al., 2014; Williamson et al., 2016). ZRS is located inside intron 5 of *Lmbr1* near CTCF3 and the *Lmbr1* TAD boundary (Figure 4.1). I performed DNA FISH assay to look at the spatial proximity of *Shh*, SBE2 and ZRS in E11.5 wild type and Δ CTCF mutant posterior (ZRS-regulated, *Shh*-expressing) and anterior (control, non-expressing) limb bud tissue (Figure 4.3).

I observed increased co-localisation frequency between *Shh* and ZRS in wild type posterior compared to anterior limb bud tissue and in line with previous observations (Figure 4.3B-D) (Williamson et al., 2016). I found that in Δ CTCF1, 2 and 3 mutant embryos the spatial distance from *Shh* to SBE2 and ZRS were significantly increased in both posterior and anterior limb tissue (Figure 4.3B & C). This increased spatial distance across both tissues suggest that the loss of CTCF binding sites impacts overall TAD structure irrespective of gene expression or enhancer activation. The proportion of co-localised *Shh*/ZRS alleles in posterior limb tissue (~35%) was reduced to those observed in non-expressing anterior tissue levels (~20%) (Figure 4.3D).

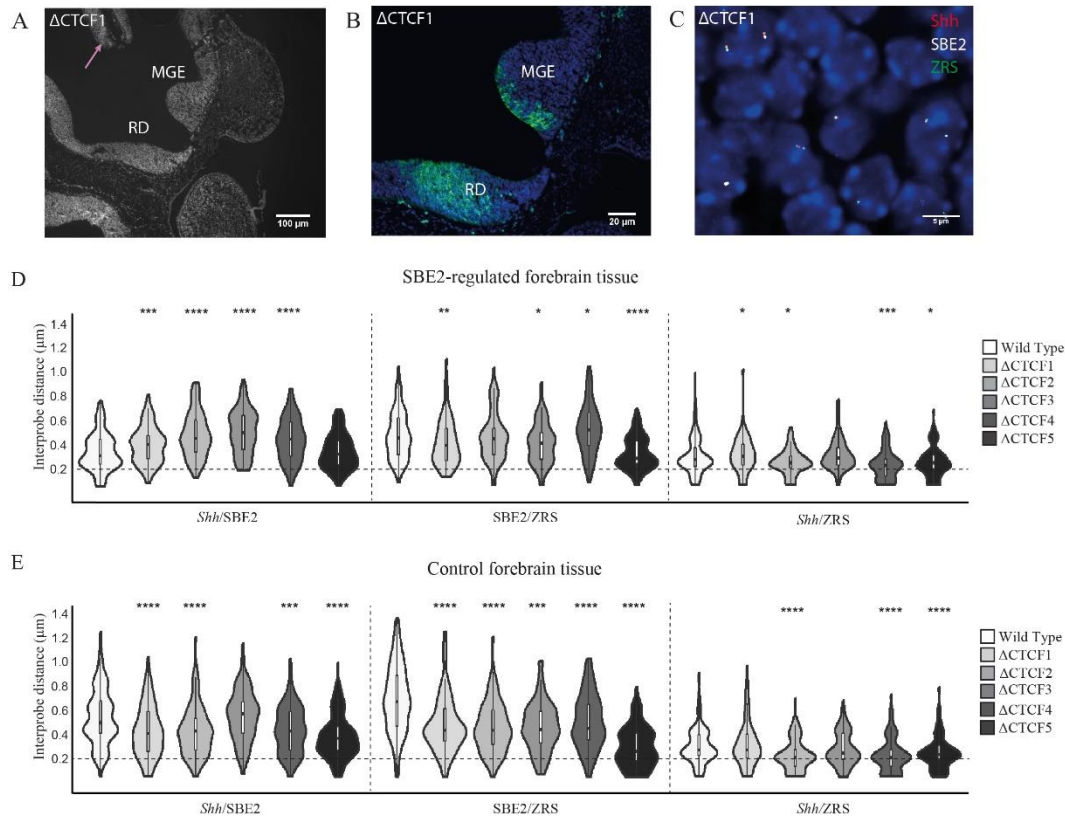


Figure 4.2 SBE2 shows increased distance from the *Shh* gene in the forebrain of Δ CTCF mutant embryos.

(A) Representative image of an E11.5 sagittal forebrain section from a Δ CTCF1 embryo showing the rostral diencephalon (RD) and medial ganglionic eminence (MGE). (B) Immunofluorescence staining of *Shh* (green) showing the SBE2-regulated region in the RD of the forebrain. (C) Images of representative nuclei from the forebrain of Δ CTCF1 embryo showing DNA FISH signals for *Shh*/SBE2/ZRS probes. (D) & (E) Violin plots show the interprobe distance (μ m) distribution between *Shh*/SBE2, SBE2/ZRS and *Shh*/ZRS probes from E11.5 wild type and Δ CTCF embryos in (D) SBE2-regulated and (E) control forebrain tissue. Horizontal dashed lines show the proportion of alleles that are colocalised (< 200 nm). The statistical significance between data sets was examined by Mann-Whitney U Tests, * < 0.05 , ** < 0.01 , *** < 0.001 , **** < 0.0001 . For each genotype there is $n = \sim 100$ nuclei from two different tissue sections from a single embryo.

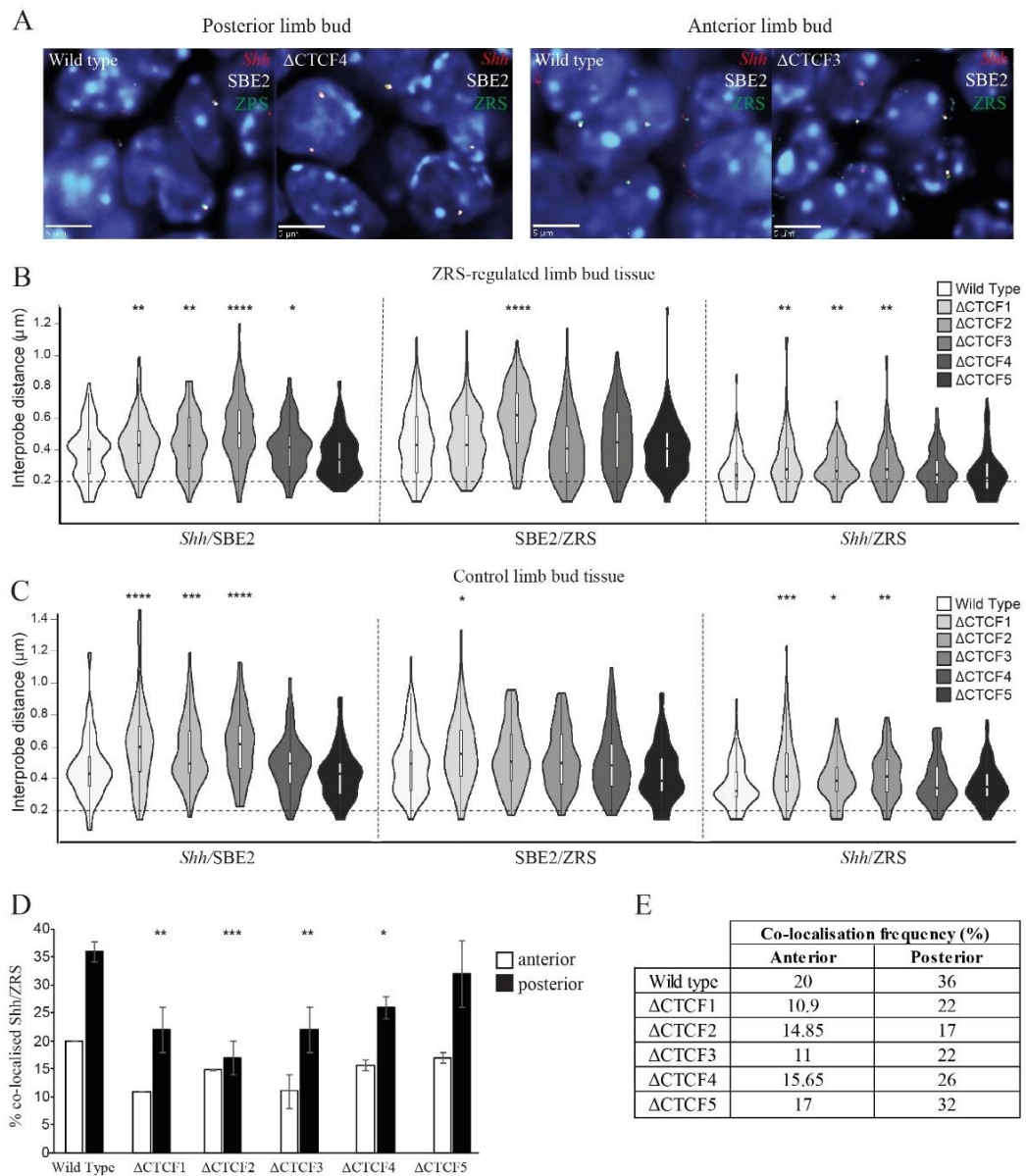


Figure 4.3 Decreased ZRS co-localisation with the *Shh* gene in the posterior limb bud of Δ CTCF mutant embryos.

(A) Images of representative nuclei from E11.5 posterior and anterior limb bud in wild type and Δ CTCF embryos showing FISH signals for *Shh*/SBE2/ZRS probes. Scale bars = 5 μ m. (B) & (C) Violin plots showing the distribution of interprobe distances (μ m) between *Shh*/SBE2, SBE2/ZRS and *Shh*/ZRS probes from E11.5 wild type and Δ CTCF embryos in (B) ZRS-regulated and (C) control limb bud tissue. Horizontal dashed lines show the proportion of alleles that are co-localised (< 200 nm). The statistical significance between data sets was examined by Mann-Whitney U Tests, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001. (D) Histograms show the percentage of co-localised *Shh*/ZRS probe pairs in wild type and each of the Δ CTCF E11.5

embryos for posterior and anterior limb bud tissue. Error bars represent SEM obtained from two or three different tissue sections. The statistical significance between data sets was examined by Fisher's Exact Tests, * < 0.05, ** < 0.01, *** < 0.001. This figure was adapted from (Williamson et al., 2019). For each genotype there is n = ~100 nuclei from two different tissue sections from a single embryo.

For both forebrain and limb *Shh*-expressing tissues loss of CTCF binding sites impacts the spatial proximity between *Shh* and its respective tissue-specific distal enhancers with increased gene-enhancer separation observed across all tissues analysed so far. However, the observed increase in distance between *Shh* and the enhancer of interest in each tissue is variable across the different Δ CTCF mutants. Δ CTCF1, 2 and 3 consistently affect chromatin conformation of the *Shh* TAD across all tissues analysed. CTCF sites 4 and 5 have minimal effects in limb tissue and CTCF4 has some effect in forebrain tissue suggesting these sites may be less involved in the *Shh* TAD structure in comparison to CTCF sites 1-3. Despite the observed changes to *Shh* TAD structure and enhancer-promoter interactions, these CTCF site deletion mice develop normally suggesting *Shh* expression is not affected or affected enough to impact tissue development. It is unclear if these changes in *Shh* TAD structure have a subtle transcriptional consequences on *Shh* expression.

4.3 *Shh* developmental expression patterns are unaffected in Δ CTCF mutants

4.3.1 Δ CTCF mutants display normal *Shh* expression patterns

In this section, whole-mount *in situ* and qPCR experiments were carried out by Laura Lettice (Hill Group). I assisted with the qPCR analysis. For more details please refer to Williamson et al., (2019) which is included in 8.2.

Models of TAD function would suggest that perturbation to the *Shh* TAD structure and enhancer-promoter interactions would disrupt developmental expression of *Shh*. However, as discussed above, Δ CTCF mice are viable, fertile, develop normally and can be maintained as homozygous lines for multiple generations without any overt phenotypic effects. To examine the broad *Shh* spatial expression patterns in Δ CTCF mutant embryos whole-mount *in situ* hybridisation (WISH) was carried out on E11.5 embryos (Williamson et al., 2019). This embryonic stage was selected for analysis as many of the *Shh* enhancers of interest are active at this developmental point. WISH is a method used to visualise spatial patterns of mRNAs in embryos using labelled antisense probes targeting the transcript of interest. As altered inter-TAD interactions were found by 5C and DNA FISH when CTCF sites were deleted in and around the *Shh* TAD (see 4.1.1), and since CTCF is a known insulator protein, the expression patterns of some of the developmental genes found in TADs neighbouring *Shh* including *En2*, *Cnpy1* and *Mnx1* were analysed (Figure 4.1). Both *En2* and *Cnpy1* are expressed at the mid-hindbrain junction (Hirate and Okamoto, 2006; Li Song and Joyner, 2000) and *Mnx1* is normally expressing in motor neurons along the neural tube (Zelenchuk and Brusés, 2011).

No difference in *Shh* expression in any of the tissues of Δ CTCF mutant embryos compared to wild type (Williamson et al., 2019). *Shh* transcripts were detected by WISH along the midline of bisected heads, in the ZPA of limb tissue and in cells of the floorplate and notochord in all embryos (Williamson et al., 2019). As CTCF is an insulator, the removal of CTCF sites may promote inter-TAD cross-talk between regulatory elements, as has been suggested by several studies (Flavahan et al., 2016; Franke et al., 2016; Kraft et al., 2019; Lupiáñez et al., 2015). Δ CTCF1 and 2 embryos

showed normal expression patterns of *En2* and *Cnpy1* at the midbrain-hindbrain junction and did not pick up any ectopic activity from the *Shh* enhancers (Williamson et al., 2019). Nor did the *Shh* gene receive any ectopic activation from *En2* or *Cnpy1* enhancers suggesting the modified *Shh* TAD boundaries in the mutant embryos are still able to provide insulation. Δ CTCF3-5 embryos have normal *Mnx1* expression in limb buds and motor neurons and did not show any ectopic expression patterns from *Shh* enhancers and vice versa indicating the *Lmbr1* TAD boundary is also intact (Williamson et al., 2019). These data suggests that the spatiotemporal specificities of *Shh*, *En2*, *Cnpy1* and *Mnx1* enhancers are maintained despite the structural changes to the *Shh* TAD and distal interactions. Quantification of *Shh* mRNA levels in dissected limb bud and entire head tissue by qPCR showed no significant changes in total mRNA levels were found in any of the Δ CTCF mutant embryos with the exception of Δ CTCF5 mutant limb bud tissue (Williamson et al., 2019).

4.3.2 *Shh* transcriptional levels are within the normal range

WISH is a useful approach to look at gene expression domains, however it is only semi-quantitative and small changes in expression will not be reliably detected using this method. *Shh* is expressed in a relatively low proportion of the cells dissected for qPCR analysis meaning the results can vary depending on the dissection and the presence of lots of non-expressing cells and variability in embryos can make it difficult to accurately compare gene expression changes. To accurately measure any changes to *Shh* expression at the single-cell level in tissues of interest in the Δ CTCF mutants I performed RNA FISH on tissue from E11.5 embryos (Figure 4.4). RNA FISH uses directly labelled oligo probes that hybridise to the nascent *Shh* transcript allowing quantification of the number of expressing alleles in each nuclei (2.8).

In SBE2-regulated forebrain tissue, approximately 60% of alleles are expressing *Shh* in wild type tissue (Figure 4.4C). There is a modest reduction in *Shh* expression in Δ CTCF mutants but this is only significant in Δ CTCF2, 3 and 5 mutants (Figure 4.4C). In posterior limb cells of the ZRS-regulated region approximately 50% of alleles are expressing *Shh* and no significant change in this frequency is detected in any of the Δ CTCF mutant embryos (Figure 4.4D).

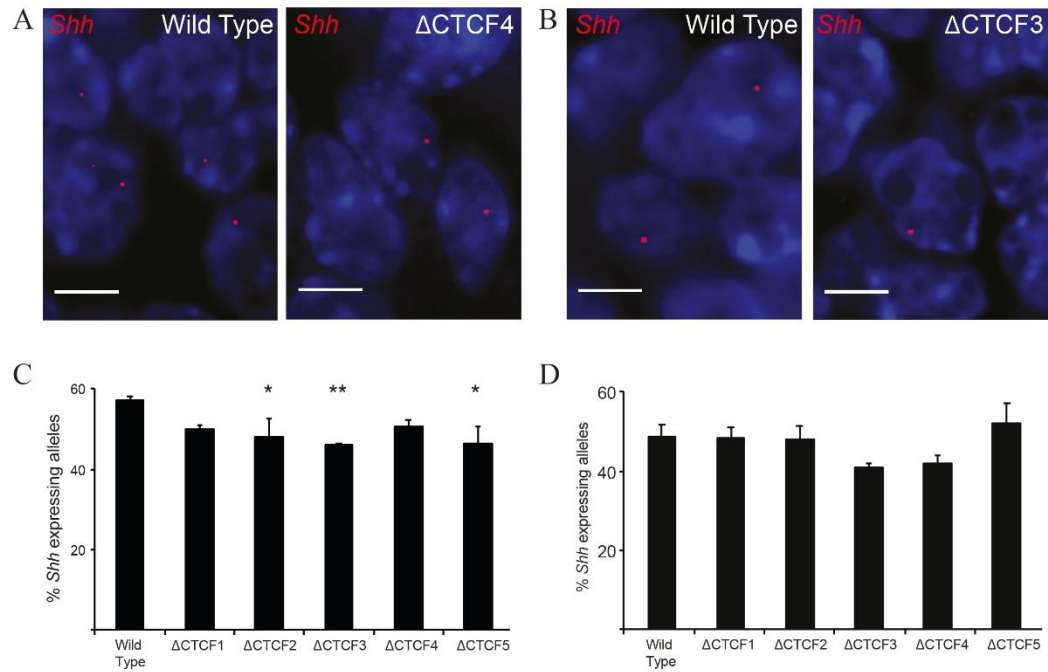


Figure 4.4 *Shh* nascent transcription is largely unperturbed following removal of CTCF sites.

(A) & (B) Images of representative RNA FISH signal (red) of *Shh*-expressing nuclei within (A) SBE2-regulated forebrain and (B) ZRS-regulated limb tissue of E11.5 wild type and ΔCTCF embryos. Scale bars = 5 μm. (C) & (D) Percentage of expressing alleles detected by RNA FISH within (C) SBE2-regulated forebrain and (D) ZRS-regulated limb tissue of E11.5 wild type and ΔCTCF embryos. The statistical significance between data sets was examined by Fisher's Exact Tests, * < 0.05, ** < 0.01. This figure was adapted from (Williamson et al., 2019). For each genotype there is n = ~100 nuclei from two different tissue sections from a single embryo.

4.4 Discussion & conclusions

In this chapter, I have presented analysis of *Shh* TAD structure, regulatory interactions and developmental gene expression from Δ CTCF mutant mESCs and embryos. 5C and DNA FISH data from Δ CTCF mutant mESCs shows that loss of CTCF binding sites changes TAD boundary positions and both inter- and intra-TAD interactions (Williamson et al., 2019). I next demonstrated using DNA FISH in Δ CTCF mutant embryonic tissues that tissue-specific distal enhancers were found at a greater distance from the *Shh* gene following CTCF site removal (Figure 4.2 & Figure 4.3). Using WISH and qPCR, it was shown that spatial expression patterns of *Shh* or neighbouring genes were normal and *Shh* mRNA levels were unaffected in Δ CTCF mutant embryos (Williamson et al., 2019). I used RNA FISH to show that the proportion of expression alleles was mostly unaffected in Δ CTCF mutant embryos with a mild reduction in the forebrain from Δ CTCF2/3/5 (Figure 4.4). Overall, I concluded that loss of single CTCF binding sites seem to consistently have consequences on *Shh* TAD structure but no overt impact on *Shh* developmental regulation.

4.4.1 Developmental gene regulation of *Shh* is not solely determined by CTCF-mediated interactions

As the single CTCF site knockouts did not affect developmental gene regulation, a larger 35 kb deletion that removes the *Lmbr1* promoter and the *Lmbr1* TAD boundary was generated (Williamson et al., 2019). Removal of this 35 kb region re-located the *Lmbr1* TAD boundary to the *Nom1* gene in the neighbouring TAD and the ZRS was found closer to the *Mnx1* gene in Δ 35 mESCs (Williamson et al., 2019). Reporter studies show that ZRS activity can extend beyond the *Lmbr1* TAD boundary (Anderson et al., 2014). *Mnx1* is weakly expressed in the limb bud cells normally but this is believed to be a by-product of strong ZRS activity in these cells rather than *Mnx1* being involved in limb development. Consequently, *Mnx1* expression is moderately upregulated in Δ 35 mutant embryos (Williamson et al., 2019). These data suggest that removal of single CTCF sites is not sufficient to disrupt inter-TAD regulatory insulation even though inter-TAD interactions were found by 5C and DNA FISH.

One explanation for this might be that *Shh* enhancers may have an inherent preference for the *Shh* promoter over others with the exception of ZRS/*Mnx1*. *Rnf32* is found within the same TAD as *Shh* and is not subject to activation by any of the *Shh* enhancers under normal conditions. In the larger 35 kb deletion multiple CTCF sites are removed and this may de-insulate or further weaken the *Lmbr1* boundary leading to ZRS-mediated *Mnx1* upregulation. In all the single CTCF site knockouts and the larger 35 kb deletion, the *Shh* TAD did not merge together with the neighbouring TADs as previously observed in other studies of structural variants removing large DNA regions encompassing TAD boundaries (Fabre et al., 2017; Lupiáñez et al., 2015). Redundancy between the CTCF sites discussed here and in other studies may prevent the merging of adjacent TADs (Paliou et al., 2019). Presumably the presence of other CTCF sites or other large bound protein complexes such as the transcriptional machinery are sufficient to block cohesin and prevent extrusion into the neighbouring TADs as reported by other studies (Busslinger et al., 2017; Davidson et al., 2016). However, removal of all the major CTCF binding sites at the *Sox9-Kcnj2* locus did result in a merging of these TADs but still without any major effects of gene expression (Despang et al., 2019).

4.4.2 Approaches to study chromatin organisation

The 5C data discussed here is from ES cells that do not express *Shh* and it cannot be ruled out that the *Shh* TAD structural changes may differ in expressing tissue. Although TADs have been suggested to be invariant across different cell and tissue types and it was shown that the *Shh* TAD does look the same in wild-type mESCs and limb tissue so it is likely that the changes observed in Δ CTCF mESCs are relevant to tissue (Williamson et al., 2016, 2019). Moreover, the increased interprobe distance between *Shh* and its distal enhancers is generally consistent across Δ CTCF mESCs and tissue (Figure 4.2 & Figure 4.3). Performing 5C assays on different *Shh*-expressing tissues from Δ CTCF mutants is limited by low numbers of expressing cells and difficulties with dissection of regions of interest from complex tissue. An alternate approach called Oligopaint which utilises oligo probes could be used to image the entire *Shh* TAD in different expressing tissues and assess any changes following removal of CTCF binding sites (Beliveau et al., 2012).

There are some discrepancies between the changes in interactions found by 5C and interprobe distances observed by DNA FISH. For example, the 5C data shows an increased interaction between the ZRS and CTCF1 whereas by DNA FISH an increased interprobe distance between fosmids encompassing these regions. Fosmids span regions larger than the 5C bins so will be picking up other interactions. Use smaller oligo probes e.g. myTAGS, Oligopaint to get better/finer resolution to detect the changes in interactions observed by 5C.

4.4.3 CTCF binding sites provide transcriptional robustness

In all the Δ CTCF mutants, no significant changes in *Shh* mRNA expression were found despite the changes to *Shh* TAD structure (Williamson et al., 2019). *Shh* is expressed in excess in mice as mice heterozygous for the *Shh* null allele are viable, develop normally and have no phenotypic effects despite a 50% reduction in *Shh* levels. In the limb bud, *Shh* levels can be reduced to ~20% of normal levels before limb bud development is severely perturbed (Lettice et al., 2017). Laura Lettice (Hill group) generated compound heterozygotes by crossing our Δ CTCF mutants onto the *Shh*^{null} sensitised background in order to exacerbate any gene expression changes or phenotypes that were not detected in previous assays. Interestingly, only Δ CTCF2/*Shh*^{null} embryos displayed a phenotype (Laura Lettice, personal communication). These embryos showed variable midline defects similar to those found in HPE patients. Δ CTCF2 cells have a disrupted structure of the sub-TAD that contains the majority of the brain enhancers that regulate *Shh* expression in midline structures. CTCF2 may be an important site for mediating interactions within this neural sub-TAD. It is unclear which other CTCF site would form this sub-TAD chromatin loop with CTCF2 but some potential peaks with the correct orientation and position have recently been detected in embryonic midbrain tissue (Paliou et al., 2019). It would be interesting to remove these CTCF sites in Δ CTCF2 cells and assess if these double CTCF knockouts can also lead to a HPE phenotype in mice.

Why is *Shh* expressed at higher levels than required for normal murine development? *Shh* is fundamental for development and loss of *Shh* is embryonic lethal. The presence of multiple CTCF binding sites may provide robustness to the system meaning loss of

one CTCF site can be compensated by another. Indeed another group showed loss of multiple CTCF sites is required to reduce *Shh* expression (Paliou et al., 2019). The authors here also crossed their double CTCF knockouts with a hypomorphic ZRS allele and showed these embryos had limb defects due to complete loss of *Shh* expression (Paliou et al., 2019). The default *Shh* TAD may function to maximise *Shh* transcriptional output and can tolerate some chromatin structure disruption e.g. single and double CTCF binding site loss.

4.4.4 Is direct interaction between enhancers and promoters required?

Inversions encompassing either of the *Shh* TAD boundaries disrupt TAD integrity and prevent ZRS from contacting/regulating the *Shh* promoter by significantly increasing the genomic distance between these elements (Symmons et al., 2016). Our lab has shown the *Shh* TAD forms a discrete compact regulatory hub that may facilitate distal enhancer contacts with the *Shh* promoter by bringing these enhancers into close spatial proximity (Williamson et al., 2016). One of the most surprising findings of this chapter was the loss of enhancer-promoter spatial proximity without any detectable effect on *Shh* expression in developing embryos (Figure 4.2 & Figure 4.3). This brings into question the importance of direct enhancer-promoter contact for developmental gene regulation. Another study from our lab showed that the SBE6 neural enhancer (~100 kb upstream of *Shh*) in fact moves further away from the *Shh* gene during NPC differentiation when this enhancer regulates *Shh* expression (Benabdallah et al., 2019). Live-cell imaging studies of the *Sox2* locus reported that enhancer-promoter proximity did not coincide with *Sox2* expression, again, suggesting that direct enhancer-promoter interactions are not required for gene induction (Alexander et al., 2019). The loss of *Shh*/ZRS co-localisation in the Δ CTCF mutant limb tissue suggests that the chromatin loop that normally forms between these elements is lost. There are still a proportion of co-localised alleles although this is similar to non-expressing tissue levels (Figure 4.3). It can be speculated that alternatively to a stable enhancer-promoter loop, *Shh* and ZRS can still interact in these tissues but the interaction may be more transient and is normally stabilised by CTCF in order to maximise *Shh* expression in posterior limb tissue. I would have expected less stabilised interactions between *Shh* and ZRS would lead to a reduction in the proportion of *Shh*-expressing alleles but I did not find this to

be the case (Figure 3.4D). Stable enhancer-promoter loops or interactions can provide robustness to the system but may not be required for the initiation of transcription. It cannot be ruled out here, or in the other studies discussed, that a single transient interaction between the enhancer and promoter is sufficient to initiate gene transcription and once this occurs the gene continues to express even in the absence of enhancer interaction. Although single CTCF site deletions disrupted TAD boundaries and chromatin contacts, in all of the Δ CTCF mutants an overall TAD structure persisted in mESCs. The structure that remains may still be sufficient to facilitate some enhancer-promoter interactions that I did not detect by DNA FISH due to limited sensitivity or cells analysed.

Chapter 5. Synthetic approach to study *Shh* regulation

5.1 Introduction

CTCF and cohesin are proposed to cooperate to produce chromatin loops and TAD structures through loop extrusion. Chromatin looping is thought to bring distal enhancers into direct contact with their target promoters and facilitate gene activation. As discussed in TADs as a functional unit of gene regulation, genome-wide depletion of CTCF, cohesin or cohesin regulators disrupts TADs with minimal effects on gene expression (Haarhuis et al., 2017; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). We and others have shown that single or double CTCF binding site knockouts can perturb TAD structure and enhancer-promoter interactions without impacting developmental gene expression (Paliou et al., 2019; Williamson et al., 2019). These data contradict the view that TADs are essential for gene regulation and enhancer function. In the CTCF mutant embryos, transient enhancer-promoter interactions may still occur as cohesin-driven chromatin extrusion is ongoing even when CTCF binding sites are removed. Previous work utilising the AID system to acutely deplete CTCF or cohesin genome-wide focused on undifferentiated, steady-state cells and perhaps gene expression has already been initiated prior to auxin addition which may explain why only modest gene expression changes are observed. As cells begin to differentiate and initiate new gene expression programmes, CTCF and cohesin may be required to facilitate enhancer-mediated gene activation from a silent state.

Targeting of the *Shh* promoter or distal enhancers with a transcription activator-like effector (TALE) fused to repeats of the small viral acidic protein VP16 (TALE-VP64) can synthetically induce *Shh* expression in mESCs (Benabdallah et al., 2019). *Shh* is normally repressed by Polycomb in mESCs and forms a PRC1-mediated loop with *En2* and *Mnx1* from the neighbouring TADs (Boyle et al., 2020). *Shh* expression was shown to be induced using TALE-VP64 proteins targeting the *Shh* promoter, SBE6 (100 kb upstream) and SBE2 (410 kb upstream) neural enhancers (Figure 1.8) (Benabdallah et al., 2019). In this chapter, I develop a system to test if distal activation of *Shh* is dependent on CTCF or cohesin by using TALE-VP64 proteins targeting *Shh* enhancers in mESCs that have acutely depleted for either of these architectural proteins.

5.2 *Shh* expression is induced by synthetic activation of distal enhancers

5.2.1 Construction of tZRS-VP64

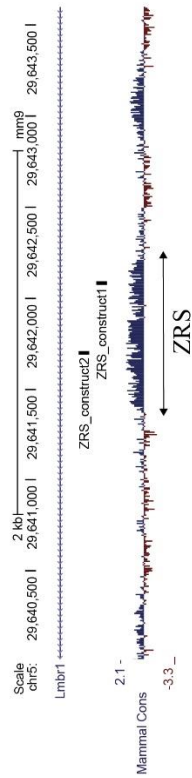
To further expand the synthetic activation system to induce *Shh* expression in mESCs from the most distally located *Shh* enhancer, I designed and constructed a TALE-VP64 to target the ZRS limb enhancer (850 kb upstream – termed tZRS-VP64).

I used ~800 bp of ZRS sequence in mouse (Ch5:29641436-29642208) to identify potential TALE DNA binding regions using TAL Effector Nucleotide Targeter 2.0 software (Figure 5.1A) (Doyle et al., 2012; <https://tale-nt.cac.cornell.edu>). I selected two target sequences due to their low G content and low number of off-target sequences elsewhere in the genome (Figure 5.1A & B). TALE-VP64 plasmids of ZRS_construct1 (L4) and ZRS_construct2 (L5) were assembled using a one-pot Golden Gate reaction (2.10.1). This method allows simultaneous assembly of multiple DNA fragments into a target vector in a desired order by utilising DNA ligase and a restriction enzyme that creates non-palindromic overhangs outside of its recognition site. As controls, I included reactions without the restriction enzyme (L1), T4 ligase (L2) or the fourth module of the DNA-binding domain (L3) (Figure 5.1C). As a positive control, I also included a reaction for a TALE-VP64 targeting the *Nrp1* promoter (L6) that had previously been designed, assembled and validated by Yatendra Kumar (Bickmore Group) (Figure 5.1C).

Following Golden-Gate assembly of reactions L1-6, I transformed competent *E. coli* cells. Controls showed as expected, a high density of colonies on spectinomycin plates with cells transformed with L1 reaction as this reaction contained the vector plasmid only (Figure 5.1C) and no colonies were found on plates transformed with reaction mixes without T4 ligase (L2) or fourth module (L3) (Figure 5.1C). Some colonies were found on the plates containing the positive control TALE-VP64 (L6) and from both of the ZRS constructs (L4 and L5) (Figure 5.1C).

I performed colony PCR using picked colonies from plates L1, L4, L5 and L6 with primers flanking the insert to screen for fully assembled DNA-binding domain of TALE-VP64. A clone containing the approximately correct band size (1650 bp) was

identified for ZRS_construct1 and sequenced to confirm that the DNA binding domain sequence was correct and did not contain any indels or single nucleotide changes (Figure 5.1D & E). I removed the VP64 activation module from the tZRS-VP64 plasmid by BamHI/BglII restriction digest to create a tZRS- Δ construct to be used as a negative control in transfection experiments as without the activation domain, *Shh* expression should not be induced. The DNA-binding domain of tZRS- Δ was also sequenced and was found to be unchanged from tZRS-VP64 (Figure 5.1E).



	Construct 1	Construct 2
RVD Sequence	HIDNGNGNINNGHIDHNINNGNINNG	HDHIDNINNGNININHDNINHDNINHG
Strand	Plus	Plus
Position	5' end of enhancer	Near middle of enhancer
Off-target counts	172	341
Target Sequence	TCTTAATCCTATAGAT	TCCAATGAACGCTCAT
Module I	CTTA	CCAA
Module II	ATCC	TGAA
Module III	TATA	CGCT
Module IV	GAT	CAT

Reaction	Contents	Expected outcome	Observed
L_1	No restriction enzyme	Colonies	+++
L_2	No T4 ligase	No colonies	
L_3	No 4th module	No colonies	
L_4	ZRS_construct1	Colonies	+
L_5	ZRS_construct2	Colonies	+
L_6	NRP1_prom	Colonies	+

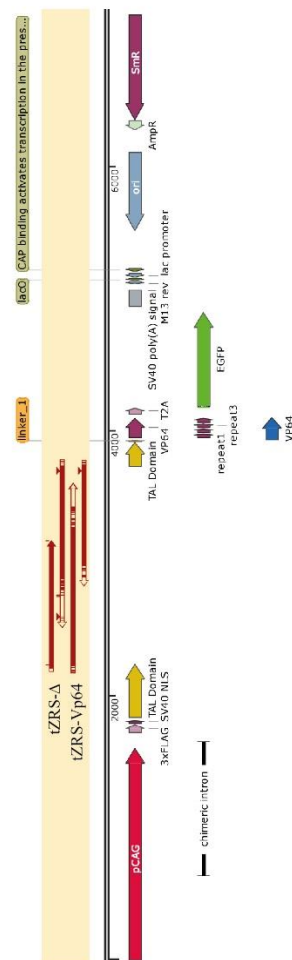
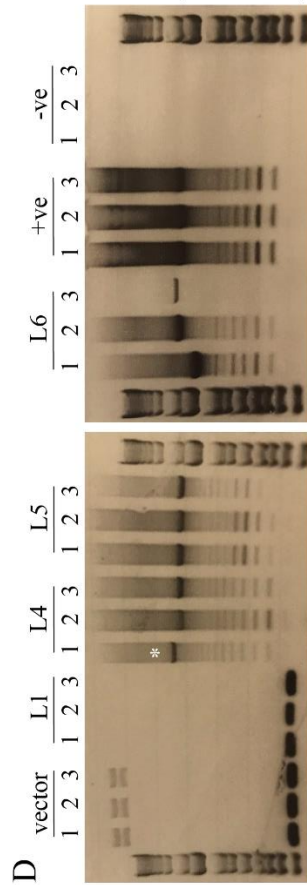


Figure 5.1 Design and assembly of a TALE-VP64 targeting the distal limb enhancer ZRS.

(A) Genomic location (mm9) of the mouse ZRS sequence showing its mammalian conservation and the positions of the potential TALE constructs (1 & 2) targeting the ZRS (black). (B) Details of the potential TALE constructs targeting the ZRS including the repeat variable diresidue (RVD) sequence of the TALE DNA-binding domain, target strand, position of the target region, number of off-target counts, target sequence of the DNA binding domain and modules used in TALE assembly. (C) Summary of the contents of the different reactions (L1-6) set up for Golden-Gate assembly and the expected and observed outcome for colonies detected following transformation. + indicates some colonies and +++ indicates many colonies. (D) Colony PCR screening to identify fully assembled TALEs. A 1650 bp band is expected for a fully assembled TALE-VP64 construct. White asterik indicates the tZRS-VP64 plasmid used for all future experiments. (E) Sequencing results from tZRS-VP64 and tZRS-Δ plasmids with reads aligned to the DNA binding domain of the TALE constructs.

5.2.2 Validation of synthetic transcription factors

I selected TALEs targeting the *Shh* promoter, SBE2 forebrain enhancer and ZRS limb enhancer for further analysis (Figure 5.3A & B). tShh is used as a positive control for induction of *Shh* expression. tSBE2 was selected as the SBE2 enhancer is found approximately in the centre of the *Shh* TAD and is not found in close spatial proximity to the *Shh* gene or any CTCF sites in mESCs. In contrast, tZRS targets the ZRS which will bind close to the *Lmbr1* boundary of the *Shh* TAD, between CTCF sites 3, and sites 4 and 5 and is found in close spatial proximity to the *Shh* gene in mESCs (Figure 5.3B). To test whether the TALE-VP64s that target distal enhancers of *Shh* induced expression, I transfected mESCs with tSBE2-VP64/-Δ and tZRS-VP64/-Δ. Cells were transfected with tShh-VP64/-Δ as a control for *Shh* induction. Non-transfected control (NTC) cells were also included in the assay. Approximately 70-80% of cells were transfected (GFP+) upon visual inspection. After 48 hours of transfection, cells were fixed and permeabilised for RNA FISH and hybridised with probes targeting the nascent transcripts of *Shh* and *Lmbr1* (Figure 5.2) and I imaged colonies that were GFP+ as this indicated the cells had taken up the TALE plasmid (Figure 5.3C, D & E). *Lmbr1* is ubiquitously expressed (Laura Lettice, personal communication) and was used as a positive control for the RNA FISH assay as the signal was not expected to be affected by the TALE transfections and TALE binding in the *Shh* TAD.

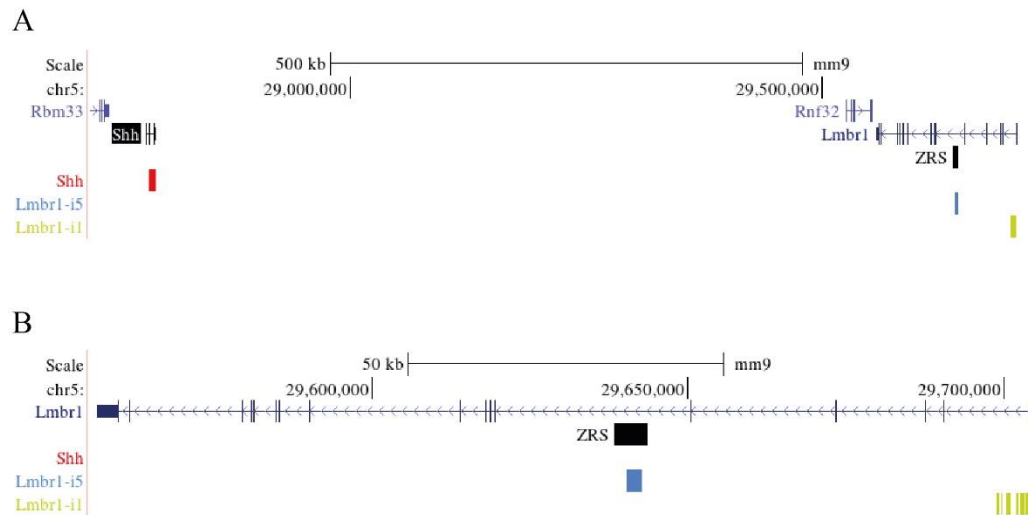


Figure 5.2 Nascent RNA FISH of *Shh* and *Lmbr1*.

(A) Schematic of the *Shh* locus (chr5:28,723,756-29,733,469) showing the position of the RNA FISH probes used to target *Shh* (red), *Lmbr1* intron 5 (*Lmbr1*-i5, blue) and *Lmbr1* intron 1 (*Lmbr1*-i1, yellow). (B) As in (A) but zoomed in around *Lmbr1* (chr5 29,555,088-29,706,234).

I initially carried out a pilot experiment and only took a small number of images in order to assess if I could detect *Shh* expression specifically in TALE-VP64 transfected wild type mESCs. I found *Shh* induction by RNA FISH in cells transfected with TALE-VP64 targeting the *Shh* promoter or distal enhancers (data not shown). However, I did not take enough images for quantification of the RNA FISH signal in this first experiment. I then repeated this assay and quantified the percentage of expressing alleles for *Shh* and *Lmbr1* signals for cells transfected with each of the TALE plasmids (Figure 5.3F). Hence data for only one biological replicate is included and I cannot say anything about levels of statistical significance. For cells transfected with tShh-VP64, I found approximately 20% of alleles expressed *Shh*. Similar levels of *Shh* expression were found in cells transfected with either tSBE2-VP64 or tZRS-VP64 (Figure 5.3F). Cells transfected with TALE-Δ showed either low level (tShh & tSBE2) or no *Shh* induction (tZRS) (Figure 5.3F). NTC cells showed no *Shh* induction (Figure 5.3F). tShh-Δ and tSBE2-Δ transfected cells had a higher level of background fluorescence than tZRS-Δ and NTC cells so some of the *Shh* signal may not represent true RNA FISH signal. For tShh and tSBE2 transfected cells, I found that

approximately 50% of alleles were expressing *Lmbr1* using probes targeting intron 5 of this transcript (Figure 5.3F). The expression of *Lmbr1* was reduced in tZRS-transfected cells. In particular, *Lmbr1* expression was reduced by around 20% in tZRS-VP64 transfected cells (Figure 5.3F). I also never detected co-expression of *Shh* and *Lmbr1* from the same chromosome in tZRS transfected cells whereas I observed co-expression to varying degrees across tShh and tSBE2 transfected cells (analysis of ~200 chromosomes per transfection). ZRS is located in intron 5 of *Lmbr1* so targeting this enhancer with a TALE seems to be interfering with the nascent transcription of this intron of *Lmbr1*. TALE binding over the ZRS may interfere with the transcriptional machinery e.g. by blocking the passage of RNA pol II. For this reason, I switched to using RNA FISH probes targeting intron 1 of *Lmbr1* (Figure 5.2). Using these probes, I found that intron 1 of *Lmbr1* is expressed from approximately 70% of alleles in cells transfected with tShh-VP64 and tShh-Δ (Figure 5.3G). In later sections (see 5.3, 5.4 & 5.5) I will go on to show that tZRS transfection does not interfere with transcription of intron 1 of *Lmbr1*. I also switched the dye used from Quasar 570 to Quasar 670 to label the *Shh* RNA FISH probes and this reduced the background fluorescence. Similar levels of *Shh* induction (~20%) are observed in tShh-VP64 transfected cells and no *Shh* expression is detected in tShh-Δ transfected cells using these *Shh* probes (Figure 5.3G). These *Shh* and *Lmbr1* RNA FISH probes are used for all subsequent RNA FISH experiments.

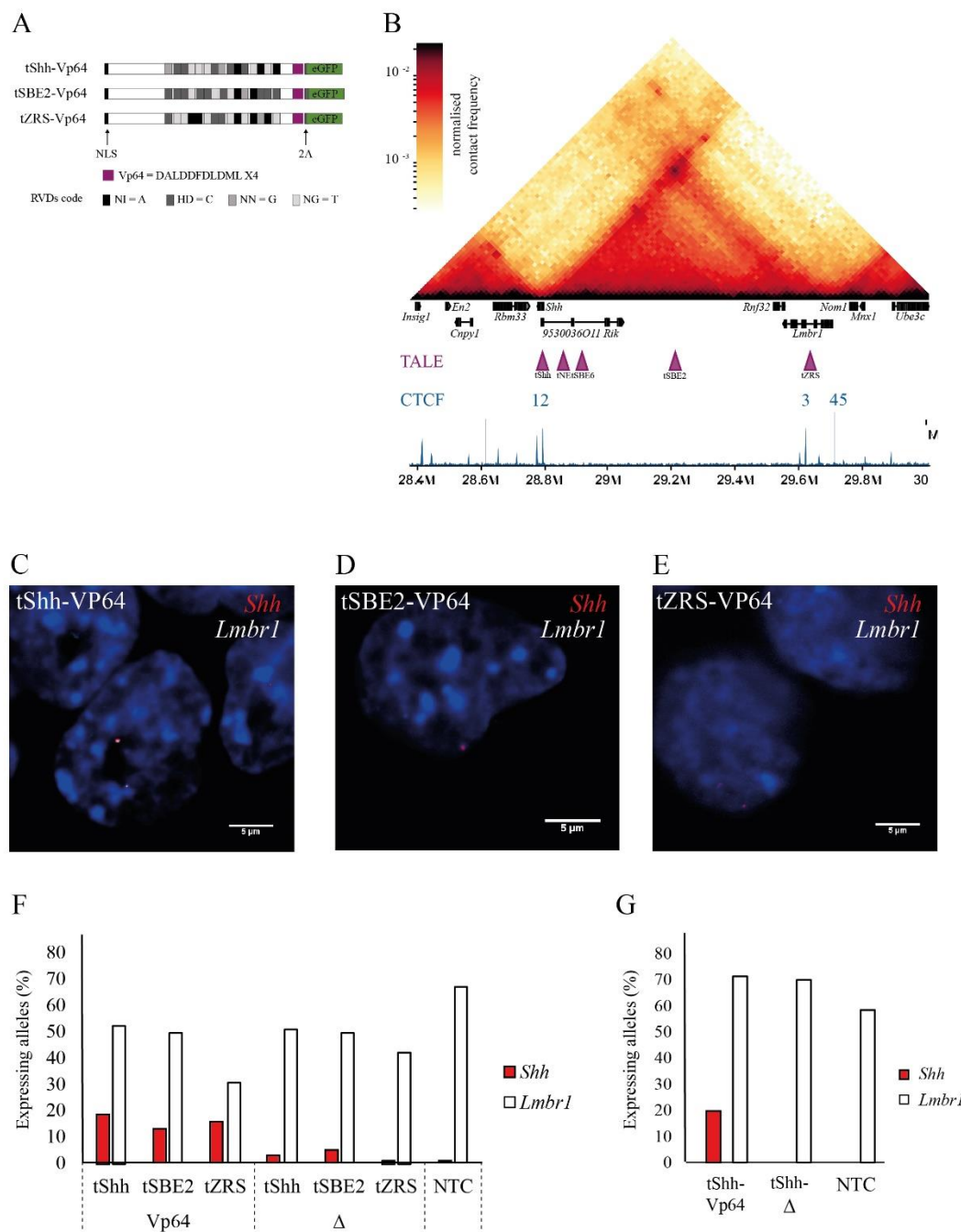


Figure 5.3 *Shh* expression can be induced using TALE-VP64 proteins targeting the *Shh* promoter or its distal enhancers.

(A) Schematic of the different TALE-VP64 constructs used to target the *Shh* promoter (tShh-VP64), non-enhancer (tNE-VP64), SBE6 (tSBE6-VP64), SBE2 (tSBE2-VP64) or ZRS enhancer (tZRS-VP64). The construct includes a nuclear localisation signal (NLS) sequence and a self-cleaving 2A peptide that allows eGFP expression. Repeat variable diresidue (RVD) code is displayed at the bottom showing the one letter

abbreviations for amino acids. **(B)** Hi-C heatmap of the *Shh* TAD from wild type mESCs at 16-kb resolution. Data shown here are from Bonev et al., 2017 and were created using HiGlass. Genes, positions of TALE target sequence and CTCF ChIP-seq track are shown below. **(C-E)** Images of representative nuclei from mESCs transfected with tShh-VP64 **(C)**, tSBE2-VP64 **(D)** and tZRS-VP64 **(E)** showing RNA FISH signals for *Shh* (red) and *Lmbr1* (white) probes. Scale bars = 5 μ m. **(F-G)** Quantification of the percentage of *Shh* and *Lmbr1* expressing alleles in TALE-transfected mESCs assayed by RNA FISH. RNA FISH probes targeting intron 5 of *Lmbr1* are used in **(F)** and probes targeting intron 1 of *Lmbr1* are used in **(G)**. A non-transfected control (NTC) is included in each experiment.

5.3 Acute degradation of architectural proteins disrupts *Shh* TAD structure

5.3.1 Degron-tagged CTCF and SCC1 are depleted upon auxin treatment

Chromatin looping between distal enhancers and promoters is mediated by CTCF and cohesin (Figure 5.4A). I obtained published cell lines where CTCF or SCC1 had been fused with an auxin-inducible degron (AID) and GFP (Figure 5.4B) (Nora et al., 2017; Rhodes et al., 2020). SCC1 (also known as RAD21) is a subunit of the cohesin complex and degradation of this component prevents cohesin from associating with DNA (Rao et al., 2017). CTCF-AID cells also contained a random integration of *Tir1* that is silenced upon cell differentiation. SCC1-AID cells contain *Tir1* integrated into the *Rosa26* locus and is stably expressed upon cell differentiation. TIR1 binds to AID-tagged proteins in the presence of auxin and targets them to the proteasome for degradation (Nishimura et al., 2009) (Figure 1.3).

To confirm that degron-tagged CTCF or SCC1 are depleted genome-wide upon auxin exposure, I treated wild type, CTCF- and SCC1-AID cells with 500 μ M auxin for 6, 24 and 48 hours. The wild type mESCs used here were obtained from Elphege Nora (UCSF) and are the parental line used to generate the CTCF-AID cell line. Untreated and treated cells were collected at the different time points and GFP fluorescence was analysed by flow cytometry to monitor protein depletion. Prior to flow cytometry, cells were also stained with DAPI to quantify the proportion of live cells.

Auxin treatment did not induce cell death in wild type or CTCF-AID cells at any time point up to 48 hours or in SCC1-AID cells after 6 hours (Figure 5.4C). 35% of SCC1-AID cells were dead following 24 hours of auxin treatment and the entire cell population was dead after 48 hours (Figure 5.4C). The cell cycle is approximately 14 hours in mESCs (Roccio et al., 2013). Cohesin is required for sister chromatid cohesion during mitosis and in its absence the SCC1-AID cells fail to divide and subsequently die. I observed that there was a high proportion of SCC1-AID cells treated with auxin that appeared stuck in mitosis following 24 hours of auxin treatment, whereas CTCF-AID auxin-treated cells appear to be able to divide in the absence of

CTCF for at least 2-3 cell cycles in line with previous reports (Nora et al., 2017). Control cells do not contain any GFP and are unaffected by auxin treatment (Figure 5.4D). FACS indicated that CTCF and SCC1 depletion occur as early as 6 hours after auxin addition and persists across all time points (Figure 5.4D). The histogram obtained for 48 hours of auxin treatment of SCC1-AID cells uses only 228 cells and the majority of the cell population is dead (Figure 5.4D). A very small proportion of GFP positive cells are found in CTCF-AID cells treated for 6 hours (Figure 5.4D). CTCF has a longer residence time on chromatin than cohesin so a longer auxin treatment may be necessary to eliminate the protein genome-wide (Hansen et al., 2017). I decided to use 24 and 6 hour auxin treatment for CTCF-AID and SCC1-AID cells, respectively, for all subsequent experiments. I performed immunofluorescence on untreated and auxin-treated CTCF-AID and SCC1-AID cells (Figure 5.4E). Very low levels of GFP signal is detected in cells treated with auxin indicating that CTCF or SCC1 protein has been depleted genome-wide (Figure 5.4E). Auxin-treated CTCF-AID and SCC1-AID cells displayed normal mESC colony formation and growth during growth conditions.

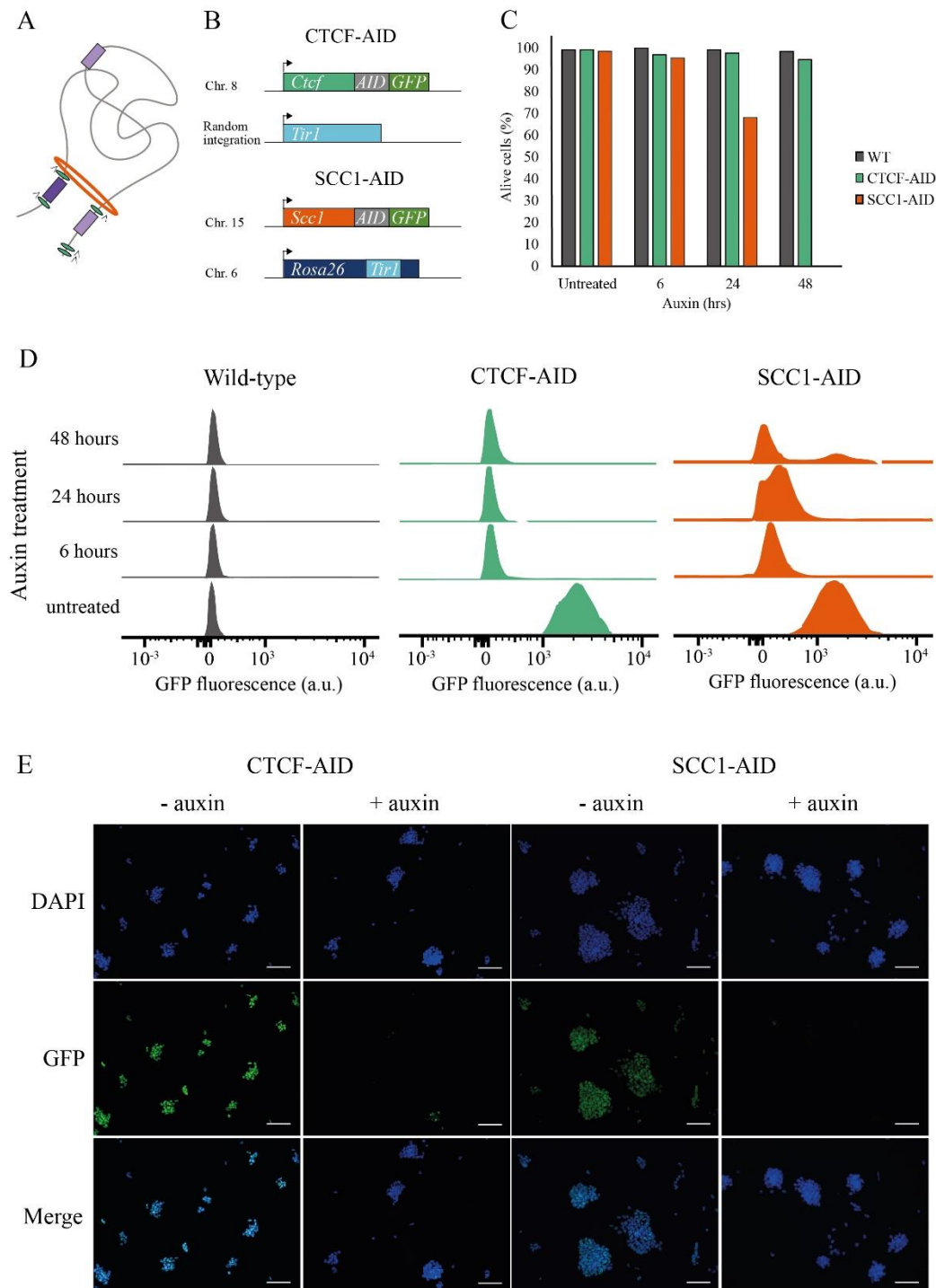


Figure 5.4 Auxin treatment leads to targeted degradation of AID-tagged proteins.

(A) Schematic of a chromatin loop showing a gene (purple) brought into close proximity to an enhancer (lilac). This interaction is proposed to be mediated by CTCF (green) and cohesin (orange). (B) Schematic illustrating the genotype of the CTCF-AID and SCC1-AID mESC lines. Endogenous *Ctcf* or *Scc1* is tagged with an auxin-inducible degron (AID) and GFP. *Tir1* is randomly integrated in the CTCF-AID line

and inserted into the *Rosa26* locus in SCC1-AID cells. (C) Quantification of live cells by DAPI staining during flow cytometry in untreated and auxin-treated wild type (WT), CTCF-AID and SCC1-AID cells after 6, 24 and 48 hours of growth in auxin. (D) Histograms of GFP fluorescence (a.u.) following flow cytometric analysis in wild type, CTCF-AID and SCC1-AID cells following 6, 24 and 48 hours of auxin treatment. N.B. in the histogram for 48 hours of auxin treatment of SCC1-AID cells the majority of cells were dead and only 228 cells are measured here. (E) Immunofluorescence staining of untreated (- auxin) and treated (+ auxin) CTCF-AID and SCC1-AID cells after 24 and 6 hours of growth in auxin, respectively. Scale bar = 100 μ M.

5.3.2 CTCF or SCC1 depletion disrupts the *Shh* TAD

Hi-C analysis showed that targeted degradation of CTCF or SCC1 removes TADs and CTCF/cohesin-mediated chromatin loops (Nora et al., 2017; Rhodes et al., 2020). I used these published Hi-C datasets to look at the impact of loss of CTCF or SCC1 on the *Shh* TAD structure. Untreated CTCF-AID and TIR1 control cell lines have a TAD structure very similar to wild type mESCs (compare Figure 5.3B and Figure 5.5 – left hand panels). The *Shh* TAD boundary is located at CTCF sites 1 and 2 and the *Lmbr1* boundary is located at sites 4 and 5. A peak interaction is found at the top of the Hi-C map formed by a chromatin loop between CTCF sites 1/2 and 3. In auxin-treated CTCF-AID cells, the *Lmbr1* TAD boundary in particular is weakened and there is more inter-TAD interactions between the *Shh* and neighbouring *Mnx1* TAD (Figure 5.5A – right hand panel). The *Shh* TAD boundary appears to partially persist in the absence of CTCF and some residual CTCF binding detected by ChIP-seq at site 1 which may account for the persistence of this boundary (Figure 5.5A). The chromatin loop between CTCF1/2 and 3 is absent in auxin-treated CTCF-AID cells. The *Shh* neural sub-TAD structure is still evident in auxin-treated CTCF-AID cells suggesting that this structure may form independently of CTCF (Figure 5.5A). In contrast, auxin treatment of SCC1-AID has a pronounced effect on *Shh* TAD structure (Figure 5.5B – right hand panel). Both *Shh* TAD boundaries are eliminated, there is no inter-TAD interaction and the majority of intra-TAD interactions are abolished (Figure 5.5B). Only strong enrichment is found along the diagonal indicating that only short range interactions persist in the absence of cohesin (Figure 5.5B). There is some interaction detected within the *Shh* neural sub-TAD structure suggesting that these interactions may form independently of both CTCF and cohesin in mESCs (Figure 5.5B).

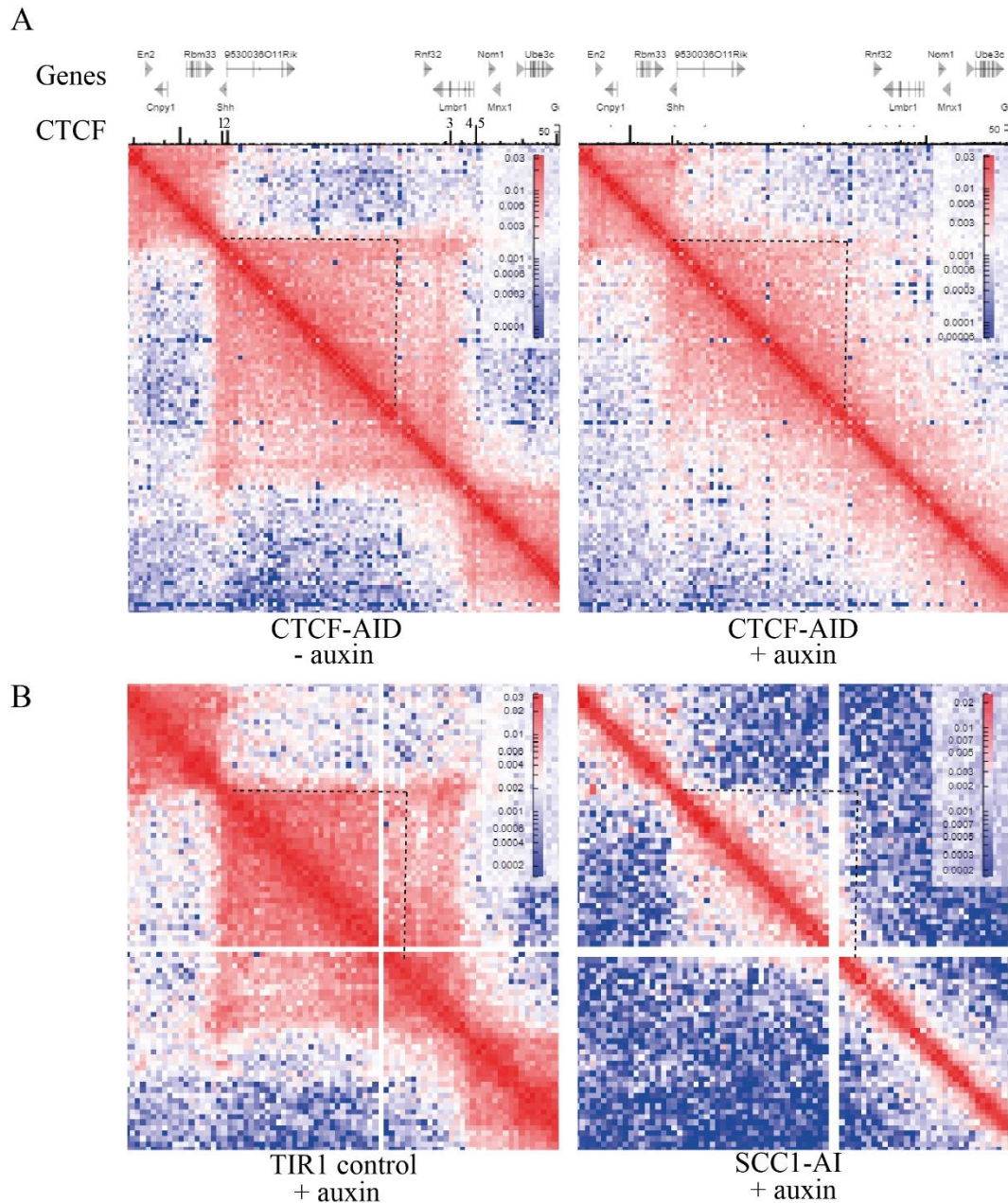


Figure 5.5 Depletion of CTCF or SCC1 disrupts the *Shh* TAD structure.

(A) Hi-C heatmaps of the *Shh* TAD from untreated (- auxin) and 48 hour treated (+ auxin) CTCF-AID mESCs at 16-kb resolution (data shown here is from Nora et al., 2017 and was created using HiGlass). Genes (grey) and CTCF ChIP-seq (black) tracks are shown above with CTCF sites 1-5 numbered. CTCF ChIP-seq data shown here is from Nora et al., 2017 and are from untreated (left) and auxin-treated (right) CTCF-AID cells. (B) Hi-C heatmaps of the *Shh* TAD from 6 hour auxin treated TIR1 control or SCC1-AID mESCs at 20-kb resolution (data shown here is from Rhodes et al., 2020 and was created using HiGlass). The dotted line marks the *Shh* neural sub-TAD.

5.3.3 Auxin treatment does not affect TALE activity in mESCs

Auxin is a plant hormone that is not normally found in mammalian cells and is an essential component of the AID system (Natsume et al., 2016). To ensure addition of auxin to the mESC culture medium did not impact the activity of the TALEs, I transfected wild type mESCs with TALE-VP64/- Δ targeting the *Shh* promoter, SBE2 and ZRS and added auxin to the media on day 2 of transfection for 24 hours. I collected, fixed, permeabilised and hybridised transfected cells with RNA FISH probes targeting *Shh* and *Lmbr1* (intron 1) (Figure 5.6A).

As shown in 5.2.2, targeting the *Shh* promoter or distal enhancers with TALE-VP64 leads to activation of *Shh* expression (Figure 5.6B & C). Targeting the same loci with TALE- Δ does not activate *Shh* expression. The proportion of expressing alleles quantified for all TALE-transfected cells is the same in the presence or absence of auxin indicating that auxin does not affect TALE activity in mESCs (Figure 5.6B & C). *Lmbr1* expression is also consistent (~60-70%) across all conditions and unaffected by the TALEs targeting the ZRS (Figure 5.6B & C). In comparison to previous experiments (see 5.2.2) the proportion of *Shh*-expressing alleles has reduced from 20% to 10% when targeting the *Shh* promoter with TALE-VP64 (Figure 5.6). Similarly, tSBE2-VP64 and tZRS-VP64 transfected cells are expressing *Shh* from 5-8% of alleles (previously 12-18% - Figure 5.6F). This reduction in expression is independent of auxin addition and consistent across two biological replicates. *Lmbr1* expression has also reduced by approximately 10% (from 70 to 60% expressing alleles) in these experiments (Figure 5.6). It is unclear why the efficiency of the TALE-VP64s has changed between these experiments. In this experiment I am using the wild type mESCs that I obtained from Elphege Nora (UCSF) that are the parental cell line of CTCF-AID cells. Previously I had used an in-house wild type mESC line but both used are E14 cells. However, despite the reduction in expression, *Shh* can still be activated when targeting the promoter region or its distal enhancers with TALE-VP64 proteins in this cell line.

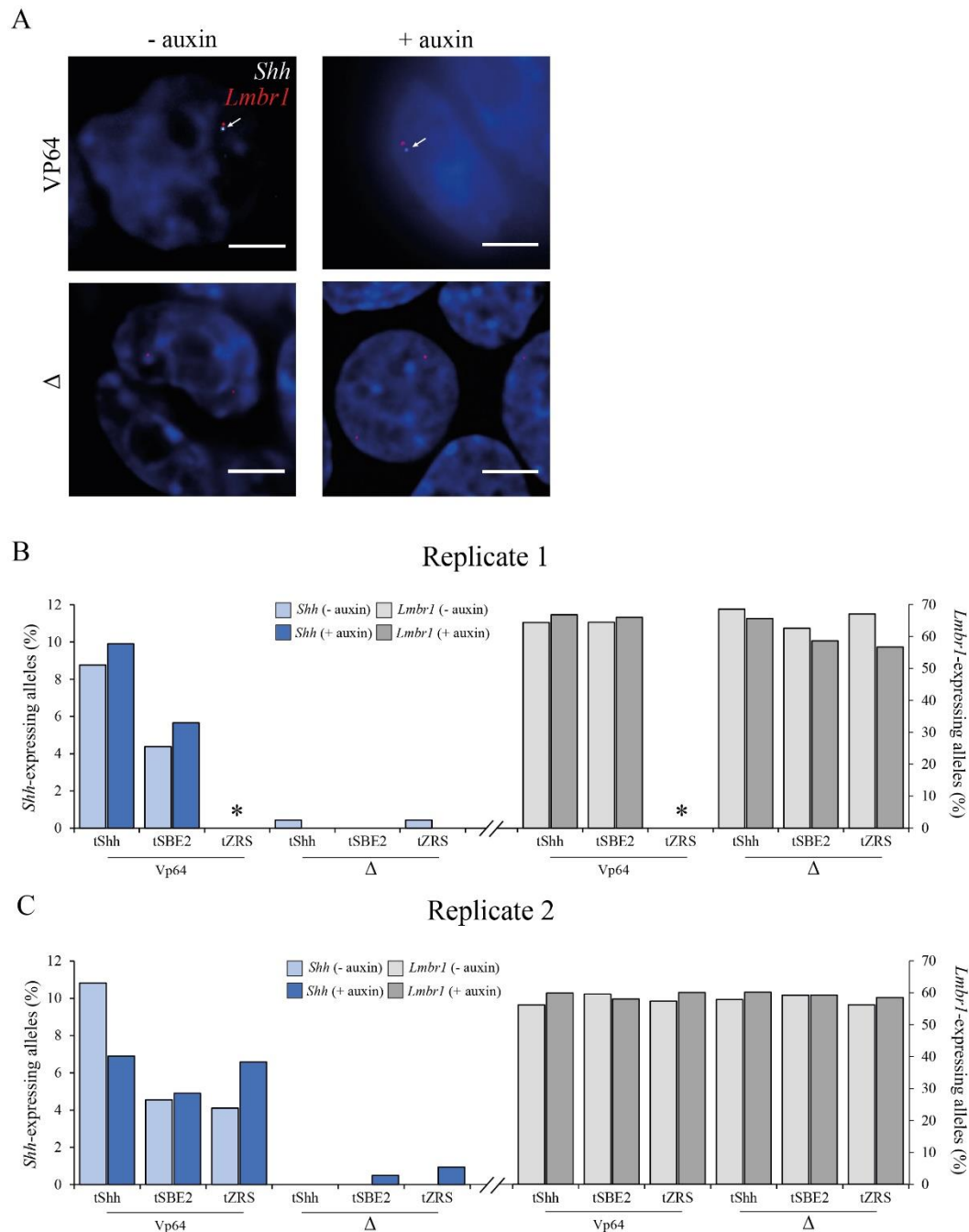


Figure 5.6 Auxin treatment does not affect TALE-mediated *Shh* induction.

(A) Images of representative nuclei showing RNA FISH signals for *Shh* (white) and *Lmbr1* (red) probes from wild type mESCs transfected with tShh-VP64/ Δ either untreated (- auxin) or treated with 24 hours of auxin (+ auxin). *Shh* RNA FISH signal is indicated by white arrow. Scale bars = 5 μ m. (B & C) Quantification of the percentage of *Shh* (blue) and *Lmbr1* (grey) expressing alleles in TALE-transfected mESCs either untreated (- auxin) or treated with 24 hours of auxin (+ auxin) assayed by RNA FISH. Data shown obtained from two independent biological replicates. Data not included in (B) is indicated by an asterisk.

5.4 Distal activation of *Shh* does not depend on CTCF

CTCF is frequently found at the base of chromatin loops and is thought to mediate long-range enhancer-promoter interactions. We have previously shown that the *Shh* TAD contains a number of CTCF binding sites that are important for TAD structure and enhancer-promoter interactions but do not individually seem necessary for developmentally-regulated *Shh* (Williamson et al., 2019). To test if loss of CTCF protein and therefore CTCF binding across sites 1-5 of the *Shh* TAD affected the ability of distal enhancers to activate *Shh* expression, I transfected CTCF-AID cells with tShh-/tSBE2- and tZRS-VP64, and with TALE-Δs as negative controls. Auxin was added to the media of cells on day two of transfection for 24 hours. I collected, fixed, permeabilised and hybridised transfected cells with RNA FISH probes targeting *Shh* and *Lmbr1* (intron 1) (Figure 5.7A) and quantified the levels (proportion of alleles) of *Shh* and *Lmbr1* expression.

In the absence of auxin, CTCF-AID cells expressed *Shh* from 10% of alleles when transfected with tShh-VP64 and from 5% of alleles when transfected with either tSBE2-VP64 or tZRS-VP64 (Figure 5.7B). Upon auxin addition, CTCF is depleted genome-wide and *Shh* TAD structure is disrupted in CTCF-AID cells (Figure 5.4 & Figure 5.5), but I found that *Shh* expression could still be induced when targeting either the *Shh* promoter or the enhancers with TALE-VP64 and the proportion of expressing alleles were not different in the presence or absence of CTCF (Figure 5.7B). No *Shh* expression was detected in CTCF-AID cells transfected with TALE-Δ and *Lmbr1* was expressed from ~60% of alleles across all conditions (Figure 5.7B). These data suggests that activation of *Shh* expression by targeting its distal enhancers with TALE-VP64 proteins does not require CTCF binding nor the *Shh* TAD structure formed by the presence of CTCF.

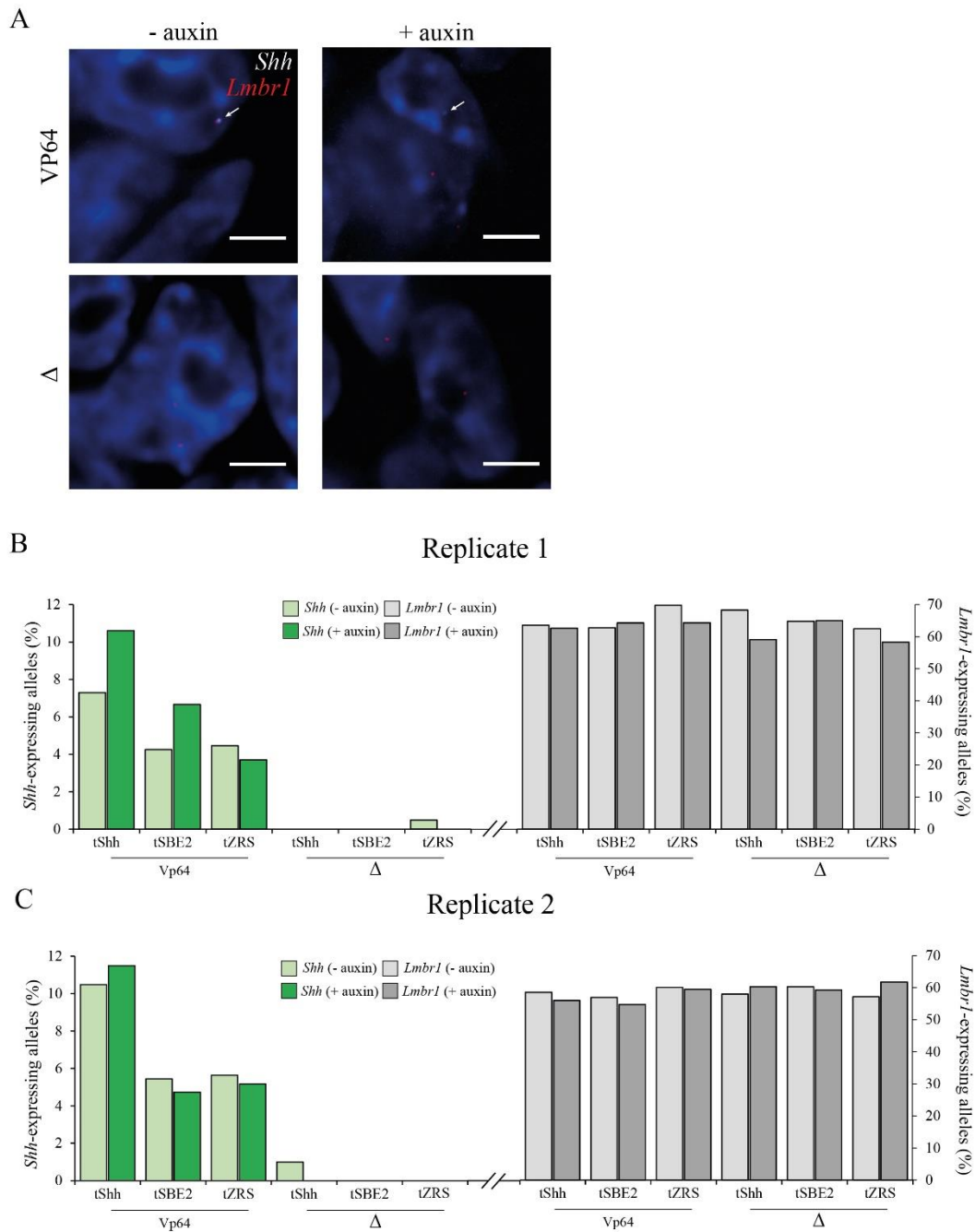


Figure 5.7 TALE-mediated *Shh* activation is independent of CTCF binding.

(A) Images of representative nuclei showing RNA FISH signals for *Shh* (white) and *Lmbr1* (red) probes from CTCF-AID mESCs transfected with tShh-VP64/ Δ either untreated (- auxin) or treated with 24 hours of auxin (+ auxin). *Shh* RNA FISH signal is indicated by white arrow. Scale bars = 5 μ m. (B & C) Quantification of the percentage of *Shh* (green) and *Lmbr1* (grey) expressing alleles in TALE-transfected CTCF-AID mESCs either untreated (- auxin) or treated with 24 hours of auxin (+ auxin) assayed by RNA FISH. Data shown obtained from two independent biological replicates.

5.5 Cohesin is necessary for distal activation of *Shh*

The cohesin complex is essential for the formation of chromatin loops and TADs and is frequently found co-localised on chromatin with CTCF by ChIP. We have not previously tested the role on cohesin on *Shh* developmental regulation as this is an essential protein and does not have site specific DNA-binding activity like CTCF and therefore DNA sequence motifs that can be mutated. To assess if the distal enhancers of *Shh* can activate expression in the absence of cohesin, I transfected SCC1-AID cells with TALE-VP64/- Δ proteins targeting the *Shh* promoter, SBE2 and ZRS. Auxin was added to the media of cells 6 hours prior to cell collection on day two of transfection. I collected, fixed, permeabilised and hybridised transfected cells with RNA FISH probes targeting *Shh* and *Lmbr1* (intron 1) (Figure 5.8A).

Similar to the results in 5.4, SCC1-AID cells expressed *Shh* from 10% of alleles when transfected with tShh-VP64, from 5% of alleles when transfected with tSBE2-VP64 and from 8% of alleles when transfected with tZRS-VP64 in the absence of auxin (Figure 5.8B). In the presence of auxin, *Shh* is still expressed from 10% of alleles when SCC1-AID cells are transfected with tShh-VP64 (Figure 5.8B). SCC1-AID cells transfected with tSBE2-VP64 or tZRS-VP64 in the presence of auxin do not activate *Shh* expression (Figure 5.8B). As before, no *Shh* expression was detected in SCC1-AID cells transfected with TALE- Δ and *Lmbr1* was expressed from ~60% of alleles in the presence or absence of auxin. This suggests that SCC1 is necessary for distal activation of *Shh* from its enhancers.

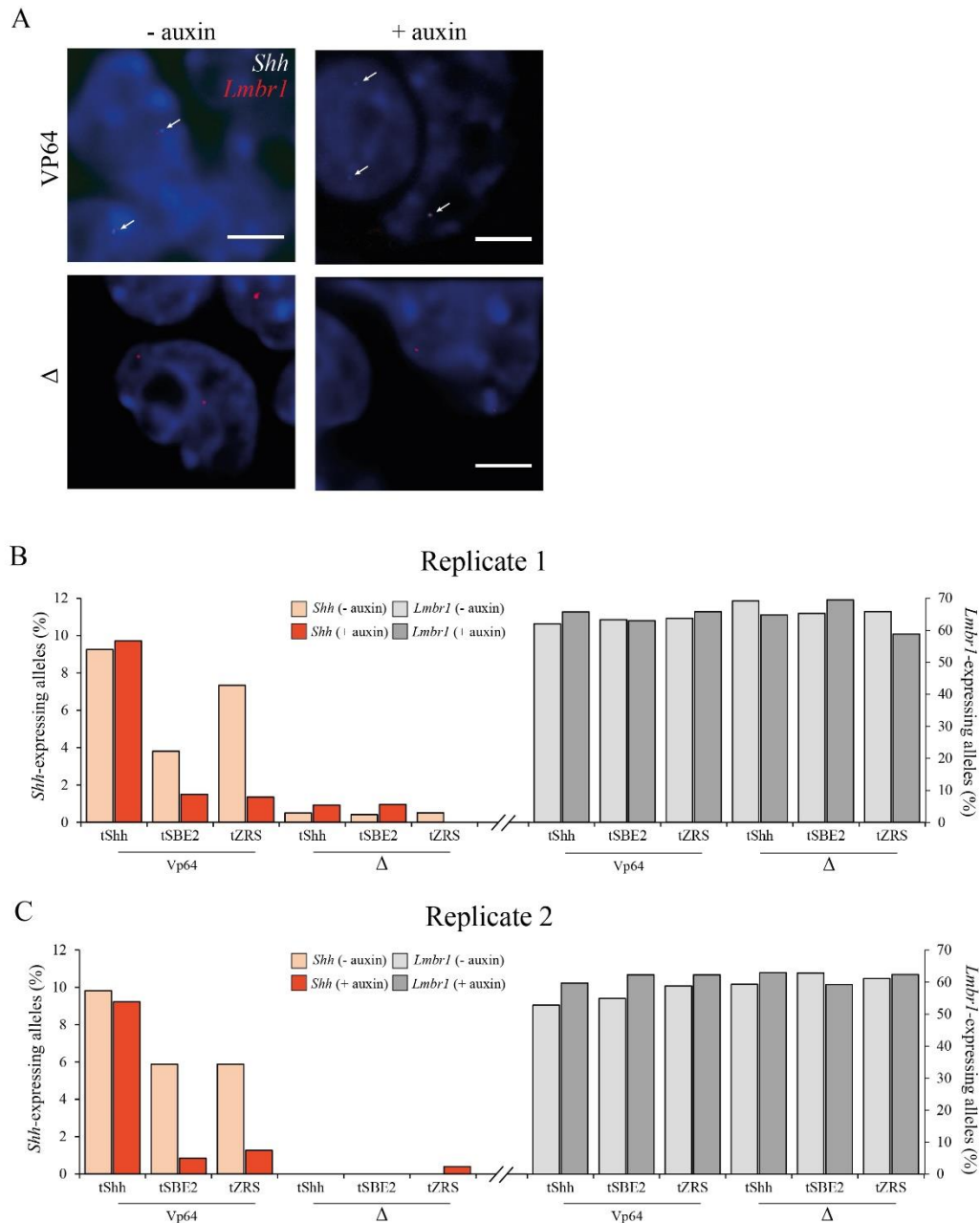


Figure 5.8 *Shh* activation using TALE-VP64 proteins targeting distal enhancers is dependent on cohesin.

(A) Images of representative nuclei showing RNA FISH signals for *Shh* (white) and *Lmbr1* (red) probes from SCC1-AID mESCs transfected with tShh-VP64/ Δ either untreated (- auxin) or treated with 6 hours of auxin (+ auxin). *Shh* RNA FISH signal is indicated by white arrow. Scale bars = 5 μ m. (B & C) Quantification of the percentage of *Shh* (orange) and *Lmbr1* (grey) expressing alleles in TALE-transfected SCC1-AID mESCs either untreated (- auxin) or treated with 24 hours of auxin (+ auxin) assayed by RNA FISH. Data shown obtained from two independent biological replicates.

5.6 Discussion & conclusions

In this chapter, I used TALE-VP64 proteins targeted to distal enhancers of *Shh* to study the relationship between chromatin structure and distal gene regulation. I show that the polycomb-mediated repression of *Shh* in mESCs can be overcome by targeting enhancers at different positions within the *Shh* TAD with TALE-VP64 proteins – including from almost 1Mb away at the ZRS. I go on to show that auxin-inducible degron-tagged cell lines can be used to deplete either CTCF or SCC1 genome-wide in mESCs and this disrupts the *Shh* TAD. Finally, I demonstrate that TALE-VP64 proteins targeting *Shh* enhancers are able to induce *Shh* expression in the absence of CTCF but are dependent on SCC1 for their activity. These findings support that CTCF functions to impede cohesin-mediated extrusion and is not directly involved in enhancer-driven gene activation. My data suggest that active extrusion of chromatin by cohesin is required to allow transient interactions to occur between distal enhancers and their target promoter leading to transcriptional activation.

5.6.1 Use of synthetic transcription factors to study gene regulation

We have previously used genetic mouse models to study the role of CTCF binding sites in *Shh* TAD structure and developmental gene regulation (Williamson et al., 2019). Mouse models can be extremely useful for studying gene regulation as phenotypes generated can provide a direct read out of the cells/tissues affected by regulatory disruptions. Generating a mouse where all CTCF binding sites of the *Shh* TAD are deleted is a time-consuming process as it requires multiple rounds of CRISPR-Cas9 genome editing. CTCF or cohesin are essential proteins so knockout model models cannot be used. Here instead, I used CTCF-AID and SCC1-AID mESC cell lines to study the effect of depletion of these proteins on gene regulation. The AID system has the advantage of allowing precise control of when a given protein is targeted for rapid degradation through the addition of auxin. This effect is also reversible by washing out auxin from cells (Nora et al., 2017). One limitation of the AID system is that tagged proteins are expressed at lower levels compared to the untagged endogenous locus (Nora et al., 2017). I also transfected a fully wild type, untagged mESC with the various TALEs in the presence or absence of auxin for comparison. No difference was found between the proportion of *Shh*-expressing alleles

in untagged and tagged cell lines transfected with tShh-VP64 (Figure 5.6Figure 5.8). This indicates that a reduction of CTCF or SCC1 protein levels in untreated tagged cell lines did not affect *Shh* induction by TALE-VP64.

One of the initial limitations of using these tagged mESC lines was that our gene of interest, *Shh*, is normally repressed in mESC. Others in the Bickmore group had previously used TALE proteins fused to a strong transcriptional activation domain (VP16/64/128) to induce expression of target genes in mESCs (Benabdallah et al., 2019; Therizols et al., 2014). *Shh* can be specifically activated using a TALE-VP64 targeted to its promoter to similar levels of mRNA expression found during neural progenitor differentiation (NPC) (Benabdallah et al., 2019). Targeting *Shh* neural enhancers with TALE-VP64 proteins can also induce *Shh* expression in mESC although less efficiently than targeting the promoter (Benabdallah et al., 2019).

In this chapter, I used the previously made tShh-VP64 and tSBE2-VP64 (Benabdallah et al., 2019), as well as generating tZRS-VP64 (Figure 5.1), to show that long-range *Shh* activation can be recapitulated by targeting distal enhancers with these proteins (Figure 5.3). In the previously published study qRT-PCR was used to measure global levels of *Shh* mRNA following transfection. I was interested to determine whether *Shh* expression could be initiated from its enhancers in the absence of CTCF or cohesin and I used an RNA FISH assay instead to quantify the proportion of expressing alleles using probes targeting the nascent transcript of *Shh*. A previous study had used PRO-seq to quantify nascent transcriptional changes following cohesin depletion with auxin and found minimal effects on gene expression (Rao et al., 2017). Cohesin loss was predicted to impact nascent transcription as genome-wide TAD/chromatin loop disruption would affect enhancer-promoter interactions. In the study by Rao et al., 2017, they focused on steady-state cells and gene expression has already been initiated prior to cohesin depletion. One benefit of using TALE-VP64 in the tagged ES cell lines is that I was able to induce expression of a gene that is normally silenced rather than studying a steady state gene. The proportion of *Shh*-expressing alleles following TALE-VP64 transfection is much lower than detected in expressing tissue (~50-60%) suggesting that *in vivo* expression is not quite recapitulated. The lack of proper

recruitment of transcription factors, co-activators etc. that are normally found at the regulatory sequences is likely a contributing factor. I also found that *Shh* and *Lmbr1* expression was reduced further by ~10% when I switched to a different wild type and tagged cell lines. This is not due to differences in transfection efficiency as only transfected cells are imaged for RNA FISH analysis. I am currently in the process of carrying out qRT-PCR analysis of *Shh* expression from these experiments to see if the differences in expression are reflected at the global mRNA level. An alternate approach to using TALE-VP64 to activate *Shh* expression would be to use the NPC differentiation (Benabdallah et al., 2016). *Shh* expression is activated during NPC differentiation by the SBE6 enhancer and it would be interesting to see if cohesin was depleted prior to *Shh* induction in this system would expression still be detectable.

5.6.2 Cohesin is required for enhancer-driven gene activation

I showed in this chapter that targeted activation of *Shh* from its distal enhancers requires the cohesin complex but not CTCF (Figure 5.7 & Figure 5.8). Initially, this result was surprising as depletion of either of these proteins has been shown to disrupt TAD structure (Nora et al., 2017; Rao et al., 2017; Rhodes et al., 2020; Wutz et al., 2017). The *Shh* TAD structure is disrupted in both auxin-treated CTCF-AID and SCC1-AID cells, albeit to differing degrees (Figure 5.5). The Hi-C experiments and analyses were performed by different labs using different tagged mESC lines may account for some of the differences observed in the Hi-C heatmaps. I am currently carrying out DNA FISH experiments in order to compare any structural differences between the *Shh* TAD following depletion of either CTCF or SCC1. DNA FISH will also be performed on cells that have been transfected with TALE-VP64 proteins targeting SBE2 or ZRS. This will provide important insights into what happens to the *Shh* TAD structure at the single cell level upon enhancer activation and if there is any difference in the absence of CTCF or SCC1. DNA FISH is limited as only a three to four positions of the *Shh* TAD can be visualised per experiment. An alternate approach such as Capture-C would provide a high resolution map of the entire TAD allowing detection of any changes chromatin interactions at any position.

SBE2 is located in the centre of the *Shh* TAD and does not appear to frequently interact with the *Shh* promoter in mESCs (Williamson et al., 2019). The neural sub-TAD structure containing the majority of the *Shh* brain enhancers persists upon CTCF/SCC1 depletion contains SBE2. ZRS is found near the *Lmbr1* TAD boundary and forms a stable chromatin loop with the *Shh* promoter in mESCs (and all tissues analysed thus far). This chromatin loop or peak interaction is lost following depletion of either CTCF or SCC1. Despite the differences in the positions and chromatin organisation of these enhancers in relation to the *Shh* promoter, both are dependent on cohesin for their activity. I speculate that the nature of an active extrusion mechanism forming chromatin loops by cohesin leads to the continuous formation of transient interactions between distant chromatin regions. If an enhancer is bound by the correct repertoire of transcription factors etc. and encounters its target promoter during extrusion then the gene will be activated. This does not necessitate the formation of a stable loop that might be detectable e.g. by FISH. A complex regulatory landscape like *Shh* may therefore rely on a general loop-extrusion mechanism for the activity of the tissue-specific enhancers. Another interesting *Shh* enhancer is SBE6 - this enhancer was found to have decreased proximity to the *Shh* promoter upon enhancer activation (Benabdallah et al., 2019). How this enhancer fits into an extrusion model of enhancer activation is unclear at this stage. I am currently performing transfections with TALE-VP64 proteins targeting SBE6 and a non-enhancer (NE) region of the *Shh* TAD in auxin-treated CTCF-AID and SCC1-AID cells to address this question. Targeting of a non-enhancer region between the *Shh* promoter and upstream distal enhancers with a TALE fused to either CTCF (tNE-CTCF) or without any fused protein (tNE-Δ) significantly reduced *Shh* expression in transfected mESCs (Benabdallah et al., 2019). It was proposed that TALE binding to this intervening region would interrupt the mechanism of distal regulation from these enhancers. I propose here that cohesin is blocked from extruding chromatin from these activated enhancers to beyond this NE site meaning that the targeted enhancers cannot activate *Shh* expression as efficiently in a cohesin-dependent manner. Here, *Shh* expression is reduced and not completely abrogated (Benabdallah et al., 2019). One explanation for this would be that not all transfected cells will be bound simultaneously by tNE and tSBE6 or tSBE2 meaning

that within the population cohesin will only be blocked by a subset of cells. This data does seem to support the view that cohesin extrudes chromatin bringing together enhancers and promoters and can be blocked by intervening proteins leading to downregulation of expression.

Previous reports indicated that cohesin depletion had modest effects on gene expression (Rao et al., 2017; Rhodes et al., 2020; Schwarzer et al., 2017). As discussed, these studies focused on steady-state expression of genes that were already transcriptionally active prior to cohesin depletion. Recently, Liu et al. (2020) used the AID system to degrade the cohesin unloader WAPL and showed that cohesin becomes stabilised on chromatin at some regions and lost at others. They found an enrichment for downregulated genes in regions depleted for cohesin. The authors propose that cohesin stabilisation reduces the pool of free cohesin in cells and disrupts loop extrusion meaning that gene expression programmes cannot be initiated (Liu et al., 2020). One study did look at the induction of the macrophage inflammatory response using an tamoxifen-inducible CRE-mediated deletion of cohesin (Cuartero et al., 2018). Genes that are normally expressed following enhancer activation during the macrophage inflammatory response are markedly downregulated in the absence of cohesin (Cuartero et al., 2018). This suggests that cohesin-mediated interactions between enhancers and promoters are needed as genes are activated in response to stimulus. Similarly during development, cohesin may function to initiate long-range enhancer-promoter interactions ensuring genes are activated during critical developmental windows. Live-cell imaging of the *Sox2* locus showed that there was no correlation between enhancer-promoter looping and transcriptional output of the gene (Alexander et al., 2019). *Sox2* is continuously expressed in mESCs and there has yet to be any live-cell imaging studies monitoring the activity of a gene as it becomes initially activated.

5.6.3 CTCF is indirectly involved in gene regulation

Single CTCF binding site knockouts affect *Shh* TAD structure but do not impact developmental gene regulation (Williamson et al., 2019). One explanation for this was that redundancy exists between CTCF sites and loss of one site can be compensated

by CTCF binding at another site. This was shown by Paliou et al., 2019, where deletion of multiple CTCF sites were required to down regulate *Shh* by ~50% but no phenotype was observed. *Shh* is expressed at greater levels than required for normal development in mouse. CTCF may function as an additional regulatory layer that increases the frequency of, or stabilises, enhancer-promoter interactions and therefore the transcriptional output. I showed that depletion of CTCF did not affect enhancer-mediated gene activation using TALE-VP64. However, this is an artificial activation system and CTCF may be required for maximising gene expression in other contexts e.g. during development. Removal of all CTCF sites surrounding the *Sox9/Kcnj9* locus in mice caused complete erasure of TAD boundaries but with no effect on gene expression (Despang et al., 2019). Inversion of the intervening TAD boundary did alter gene expression patterns and mice displayed developmental defects (Despang et al., 2019). These inversions separated enhancers from their cognate promoter leading to gene dysregulation. This data supports the model that cohesin-mediated extrusion directs enhancers to their target promoter and when an extrusion barrier is placed between these regulatory elements no transient interactions can occur.

One prediction from this model of gene regulation would be that in the absence of CTCF binding at TAD boundaries, extrusion would continue beyond TAD boundaries bringing enhancers into transient contact with non-target promoters. I am currently carrying out RT-qPCR for genes located in the TADs neighbouring *Shh* following degradation of CTCF in TALE-VP64 transfected cells to test if we see ectopic gene activation. We previously did not see any ectopic activation of *Shh* by enhancers in the neighbouring TADs following the removal of single CTCF binding sites (Williamson et al., 2019). CTCFs role may be two-fold (1) stabilise/increase interactions and (2) insulate enhancers by preventing transient interactions with non-target genes created the process of extrusion.

Chapter 6. *SHH* TAD structure in humans

6.1 Introduction

6.1.1 Acheiropodia causes severe limb defects

Acheiropodia is a rare autosomal recessive genetic disorder (~1/250,000 births) characterised by bilateral, congenital truncations of the upper and lower limb extremities i.e. loss of forearms, hands and feet (Freire-Maia et al., 1975). This disorder phenocopies the limb defects observed in humans and mice with homozygous deletions of ZRS (Chiang et al., 1996; Lettice et al., 2014; Sagai et al., 2005; Shamseldin et al., 2016). The majority of acheiropodia cases that have been identified so far have been found in consanguineous families from Brazil with the exception of a pair of siblings from Puerto Rico (Freire-Maia et al., 1975; Kruger and Kumar, 1994). The causative locus for acheiropodia was mapped to a region approximately 30 kb centromeric to the ZRS (Ianakiev et al., 2001). The *Shh*-associated limb abnormalities (1.3.6), proximity of the causative locus to the ZRS and the similarities between acheiropodia patients and the ZRS knockout phenotype led to the proposal that acheiropodia is caused by disruption to ZRS-mediated *SHH* limb regulation. However the disease mechanism of this disorder is not fully understood.

The findings discussed in this chapter are part of a collaboration between Laura Lettice (Hill Group) and Nadav Ahituv (University of California, San Francisco). Some of the findings of this chapter were included in a manuscript currently under revision (Ushiki et al., Nat. Comms.).

6.1.2 Acheiropodia patients have a Δ 12 kb deletion that encompasses CTCF binding sites

Work carried out by Laura Lettice (Hill Group) and Nadav Ahituv (UCSF) used whole-genome sequencing to further fine map the acheiropodia-associated locus to a 12 kb region in exon 4 of *LMBR1* (chr7:156608724-156620764). This region is centromeric to the human ZRS limb enhancer and does not affect the ZRS itself (Figure 6.1A). They showed using mouse transgenic enhancer reporter assays that this region does not have any enhancer activity and mice homozygous for deletion of the equivalent 12 kb region do not have an acheiropodia phenotype or any limb defects at all. Further analysis of the acheiropodia-associated region using ENCODE ChIP-seq

tracks led to the identification of three human-specific CTCF sites in this region (Figure 6.1A). No other histone marks or chromatin features were found in the 12 kb region. The CTCF sites of interest and discussed in more detail below were termed *LMBR1-SHH* CTCF (LSC) sites (Figure 6.1A).

6.1.3 Comparison of the human *SHH* TAD to mouse

I examined some published Hi-C datasets from human cell lines of the *SHH* TAD and generated a Hi-C heatmap using lymphoblastoid cell line (GM12878) data from Rao et al., 2014 (Figure 6.1B). The human *SHH* TAD is 1 Mb in size and approximately 0.1 Mb larger than the mouse *Shh* TAD (0.92 Mb). There is some overlap in the distribution of CTCF sites between the species (Figure 6.2). For details about the different CTCF binding sites in mouse and humans and their associated nomenclature please refer to Figure 6.3. CTCF1 and CTCF2 flank the *Shh* gene in mouse and similarly in humans, there are CTCF sites surrounding the gene (Figure 6.1 & Figure 6.2). A CTCF site is also detected within the human *SHH* gene body (Figure 6.1A). The TAD boundary at the *SHH* gene appears to be in the same relative location across the species and delineated by CTCF sites found at comparable positions (Figure 6.1B).

Similar to mouse, a strong peak of interaction is found at the top of the Hi-C TAD. In mouse, this peak interaction is a chromatin loop which forms between CTCF sites 1, 2 (sites which flank the *Shh* gene) and 3 (located in intron 5 of *Lmbr1*) (Figure 6.2A). Murine CTCF3 is the homologous site to LSC2 in humans (Figure 6.2 & Figure 6.3). In humans, this interaction or chromatin loop appears to be formed between convergently orientated CTCF sites found at the *SHH* gene and LSC sites 3-5 found in intron 4 of *LMBR1* (Figure 6.1B). Although this chromatin loop is present in both species, the CTCF sites found at the base of the loop at the *LMBR1* end, and presumably blocking cohesin extrusion, are found at a different positions despite the equivalent mouse site existing in human (LSC2) (Figure 6.2). The *LMBR1* boundary is difficult to define in comparison to the *SHH* boundary. In mouse, the *Lmbr1* TAD boundary is found near the TSS of *Lmbr1* and is associated with CTCF sites 4 and 5 (Figure 6.2A). There are similar CTCF sites found at the TSS of *LMBR1* in humans which may define the *LMBR1* TAD boundary (Figure 6.1A). Boundary calling

software could be used to define this but can vary depending on the algorithm used, parameters inputted etc. One striking difference between the species is the greater number of CTCF sites detected by ChIP-seq in humans compared to mouse (Figure 1.1A). The *SHH* gene desert region in humans in particular appears to contain many more CTCF sites (Figure 6.1A & 8.2). However, different antibodies are used for ChIP-seq experiments in the different species. The human *SHH* TAD has a strong sub-TAD loop that appears to be mediated by CTCF sites at the *SHH* promoter and LSC1 (Figure 6.1B). A sub-TAD containing the majority of the neural enhancers is also found in the mouse *Shh* TAD (Figure 1.8) but does not appear to have a strong CTCF loop/ peak interaction as observed for the human locus.

The 12kb acheiropodia-associated deletion removes LSC sites 3-5 that are found at the base of the strong peak interaction observed in the Hi-C map (Figure 6.1B). This led to the hypothesis that loss of LSC3-5 and disruption to the chromatin organisation of the *SHH* TAD in acheiropodia patients prevents ZRS from regulating the *SHH* promoter during limb development leading to loss of limb structures.

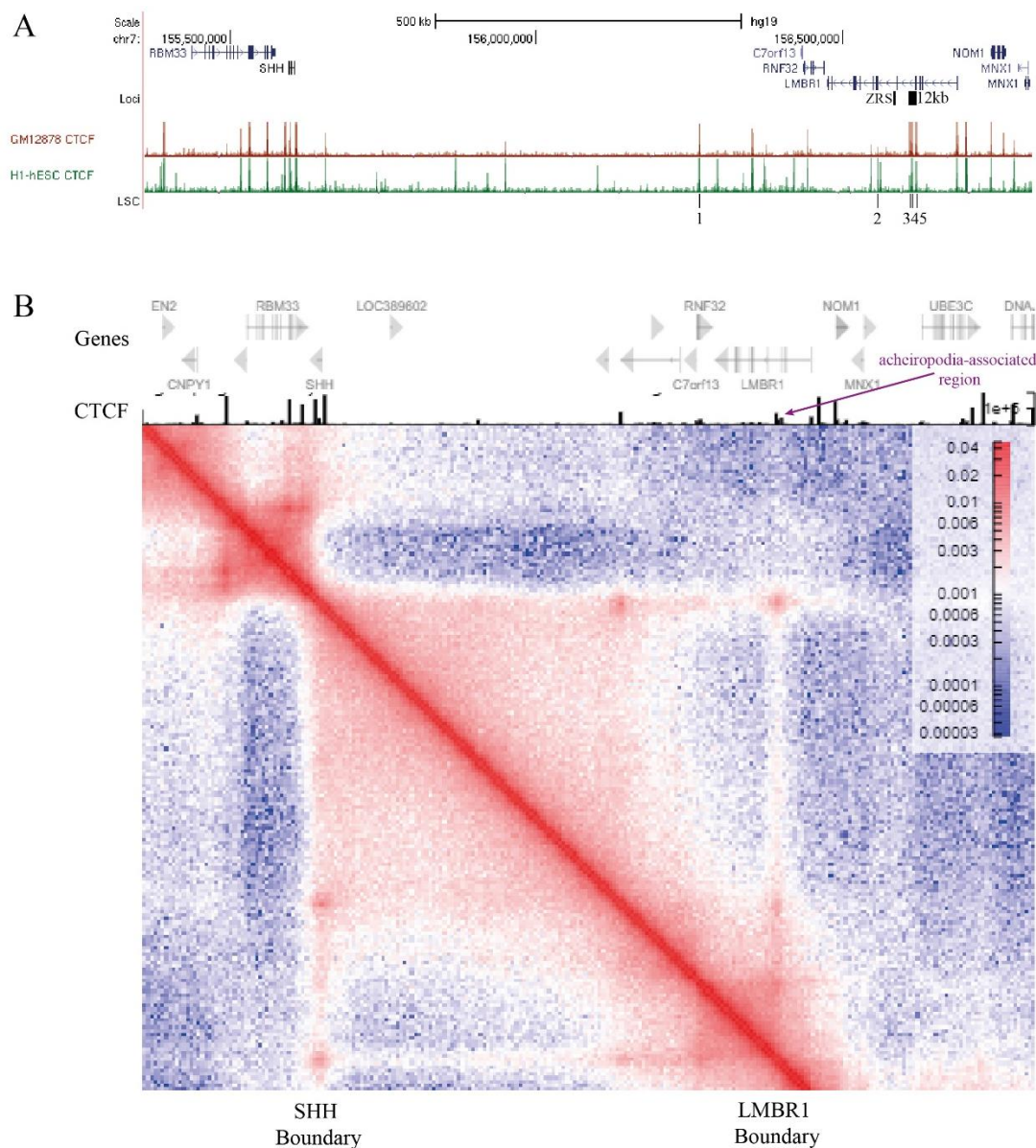


Figure 6.1 Chromatin organisation of the human *SHH* TAD.

(A) Schematic of the human *SHH* TAD and surrounding genomic regions (hg19, chr7:155,360,359-156,809,258) showing the location of genes, the human ZRS limb enhancer and the 12kb acheiropodia-associated region. The bottom two tracks show UCSC ENCODE CTCF ChIP-seq profiles from human lymphoblastoid (GM12878, red) and embryonic stem cell (H1-hESC, green) lines. The *LMBR1-SHH* CTCF (LSC) sites are numbered below (1-5). (B) Hi-C heatmap of the *SHH* TAD from wild type GM12878 lymphoblastoid cells at 8-kb resolution (data shown here is from Rao et al., 2014 and was created using HiGlass). Genes (grey) and CTCF ChIP-seq (black) track are shown above. The acheiropodia-associated region is indicated in purple. The *SHH* and *LMBR1* boundaries of the *SHH* TAD are indicated below the Hi-C heatmap.

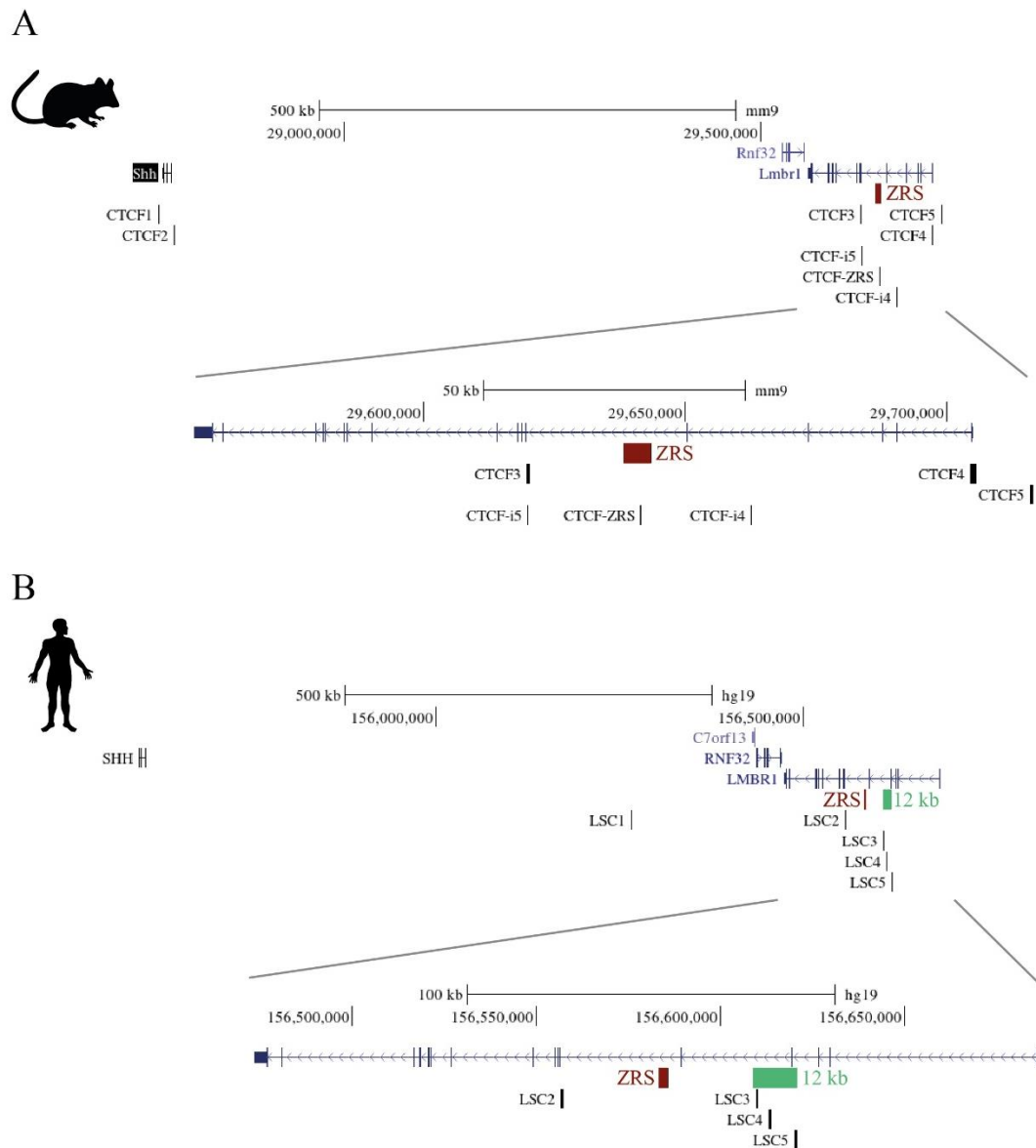


Figure 6.2 Comparison of the *Shh* TAD in mouse and humans.

(A) Schematic of the mouse *Shh* TAD (chr5:28,709,733-29,729,426) showing the location of genes, the ZRS limb enhancer (red), CTCF binding sites and a zoomed in view of the region encompassing *Lmbr1* (chr5:29,555,059-29,717,637). (B) Schematic of the human *SHH* TAD (chr7:155,546,464-156,705,086) showing gene locations, the ZRS limb enhancer (red), 12 kb acheiropodia-associated region (green), CTCF binding sites (termed LSC) and a zoomed in view of the *LMBR1* region (chr7:156,465,633-156,696,682).

CTCF binding site	Species	Genome coordinates (mm9 or hg19)	Reference
CTCF1	Mouse	chr5:28777560-28778320	(Williamson et al., 2019)
CTCF2	Mouse	chr5:28795867-28796937	(Williamson et al., 2019)
CTCF3	Mouse	chr5:29619575-29620326	(Williamson et al., 2019)
CTCF4	Mouse	chr5:29704457-29705661	(Williamson et al., 2019)
CTCF5	Mouse	chr5:29715776-29716437	(Williamson et al., 2019)
CTCF-i4	Mouse	chr5:29662545-29662612	(Paliou et al., 2019)
CTCF-i5	Mouse	chr5:29619921-29619969	(Paliou et al., 2019)
CTCF-ZRS	Mouse	chr5:29641336-29641385	(Paliou et al., 2019)
LSC1	Human	chr7:156266047-156266716	(Ushiki et al., 2021)
LSC2	Human	chr7:156556580-156557319	(Ushiki et al., 2021)
LSC3	Human	chr7:156609512-156610181	(Ushiki et al., 2021)
LSC4	Human	chr7:156613060-156613729	(Ushiki et al., 2021)
LSC5	Human	chr7:156620116-156620785	(Ushiki et al., 2021)

Figure 6.3 Summary of mouse and human CTCF binding sites at the *Shh* locus.

Table summarising the different CTCF binding sites found across the *Shh* TAD and discussed in this thesis including which species they are found, their genomic coordinates and their respective references.

6.2 Analysis of *SHH* TAD structure in human cell lines

6.2.1 DNA FISH of acheiropodia patient-derived cell lines

To assess the effect of the $\Delta 12$ kb mutation on *SHH* TAD structure, I performed DNA FISH on lymphoblastoid cell lines derived from an acheiropodia patient (MPR1) and her unaffected parents (MPR2 and MPR3). The MPR1 cell line is homozygous for the $\Delta 12$ kb mutation whereas the MPR2/3 cell lines are heterozygous. CTCF/RAD21 ChIP-seq carried out by Nadav Ahituv's group on wild type and MPR1 lymphoblastoid cell lines found that the 12 kb region is normally bound by CTCF and cohesin at LSC sites 3, 4 and 5. In the mutant cells, the loss of this 12 kb region leads to ectopic CTCF and cohesin binding at convergently orientated LSC2/CTCF3. Their data also indicated that there was increased CTCF and cohesin binding at LSC1. For DNA FISH, I used probes that hybridised to the *SHH* promoter, LSC1, LSC2 and LSC3-5 (Figure 6.4A & B).

I found that the interprobe distances between hSHH and hLSC1 were significantly decreased in patient-derived MPR1 cells compared to the MPR3 parental cell lines (Figure 6.4C & F). The DNA FISH did not work here for MPR2 cells hence why the data is not included. There was no significant difference in the interprobe distances between hSHH and hLSC2 in MPR1 cells compared to either of the parental cell lines (Figure 6.4D & F). I found that there was a significant increase in interprobe distances between hSHH and hLSC3-5, the probe which spans the 12 kb deletion, in MPR1 cells compared to the parental cell lines (Figure 6.4E & F). This suggests that the CTCF loop that normally mediates the interaction between *SHH* and LSC3-5 is perturbed in the patient cells. Instead, there appears to be an increased interaction between *SHH* and LSC1 in the patient cells suggesting this CTCF site can compensate for the loss of the three found at the acheiropodia-associated region. The ectopic CTCF/cohesin binding at LSC2 detected by ChIP-seq does not appear to modulate the interactions between this region and the *SHH* promoter in these cells.

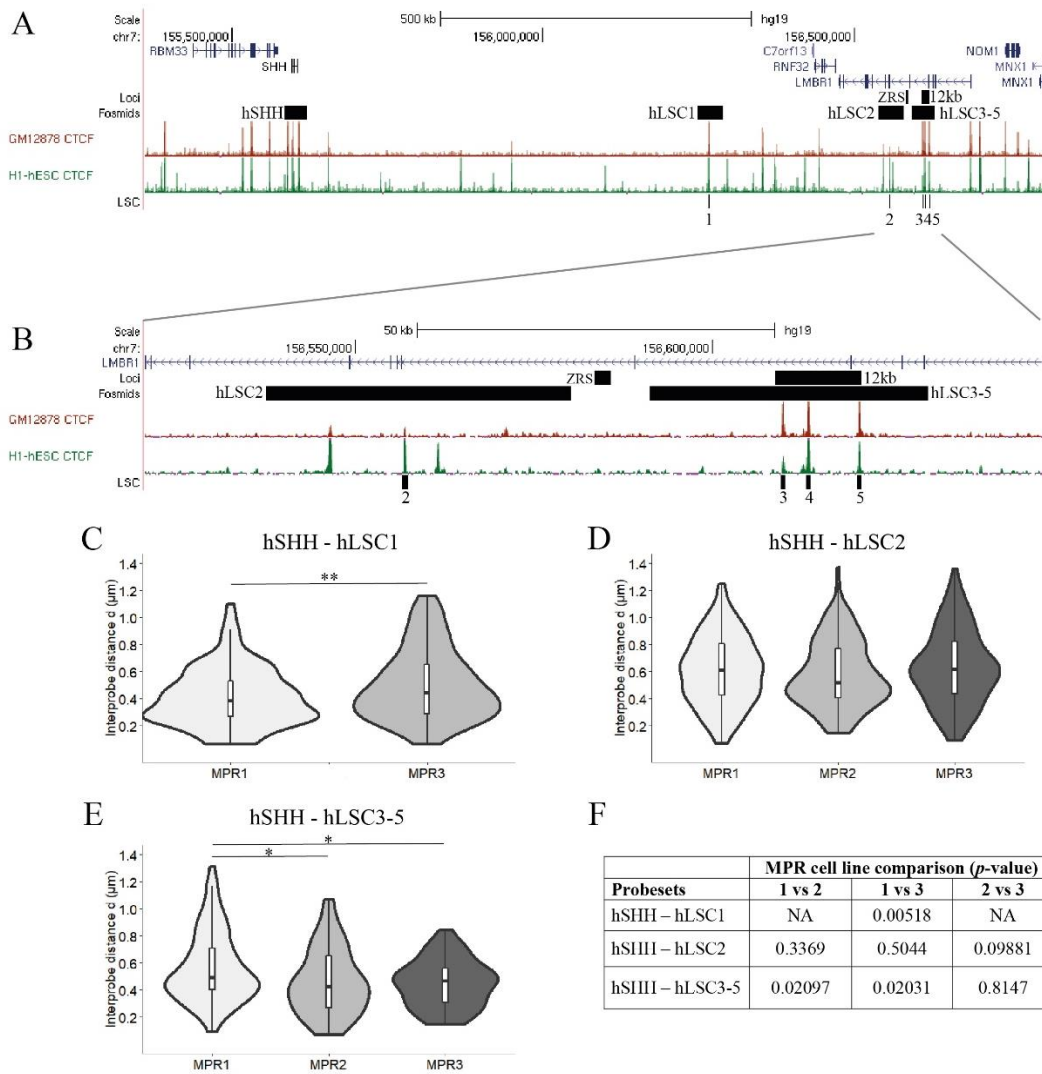


Figure 6.4 Removal of 12 kb region of the *SHH* TAD alters CTCF looping interactions in patient-derived lymphoblastoid cell lines.

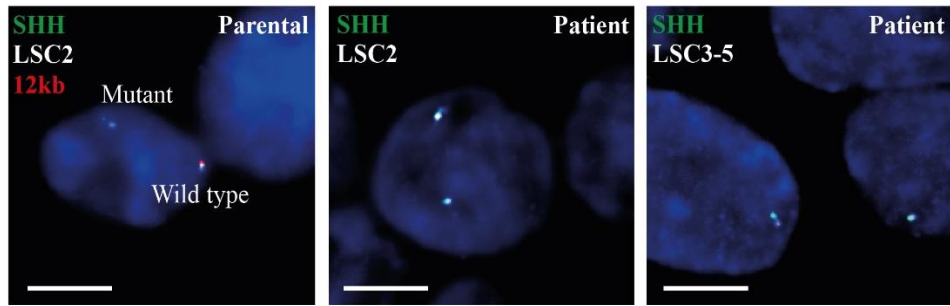
(A) Schematic of the human *SHH* TAD and surrounding genomic regions (chr7:155,360,359-156,809,258) showing the location of genes, the human ZRS limb enhancer, 12kb region lost in acheiropodia patients and location of the fosmid FISH probes. The bottom two tracks show UCSC ENCODE CTCF ChIP-seq profiles from a human lymphoblastoid (GM12878, red) and embryonic stem cell (H1-hESC, green) line. The *LMBR1*-*SHH* CTCF (LSC) sites are numbered below (1-5). (B) As in (A), zoom in of the region at *LMBR1* (chr7:156,517,979-156,644,081) (C-E) Violin plots show the distribution of interprobe distances (μm) between hSHH-hLSC1 (C), hSHH-hLSC2 (D) and hSHH-hLSC3-5 (E) probes following DNA FISH of lymphoblastoid cell lines MPR1 (patient), MPR2 (parental) and MPR3 (parental). The statistical significance between data sets was examined by Mann-Whitney U Tests, * < 0.05, ** < 0.01. $n = 75$ -150 alleles. (F) Table summarising the results (*p*-value) of Mann-Whitney U Tests comparing the interprobe distances between cell lines.

6.2.2 Allele-specific DNA FISH of patient and parental cell lines

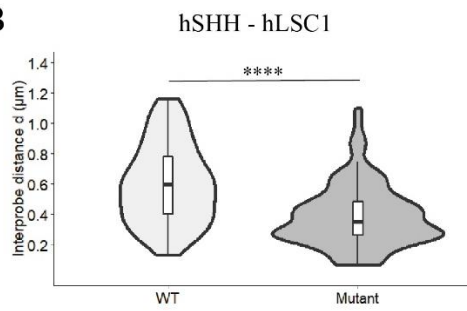
The MPR2 and MPR3 cell lines are heterozygous for the 12 kb deletion. This meant that the initial DNA FISH analysis contained a mixed population of both wild type and mutant chromosomes. A proportion of the FISH measurements will be from mutant chromosomes when ideally I would like to compare the mutant $\Delta 12$ kb allele to the wild type. The different cell lines used here therefore provided me with an opportunity to compare chromatin organisation due to sequence differences inherent to each chromosome within the same nuclei. To distinguish the two chromosomes, I used a plasmid containing the 12 kb acheiropodia-associated region (obtained from Nadav Ahituv, UCSF). I included this 12 kb probe in the DNA FISH experiments which distinguished between the wild type and mutant alleles in the heterozygous parental cell lines (Figure 6.5A). As a control, I also included this probe in DNA FISH experiments using the MPR1 cell line and never detected any signal as there is no 12 kb region for probe to hybridise to in the patient's cells (Figure 6.5A).

Similar to the results I found in section 6.2.1, the interprobe distances between hSHH and hLSC1 was significantly decreased on the mutant allele compared to the wild type (Figure 6.5B & E). An increased co-localisation frequency (<200 nm inter-probe distance) between these probes can also be detected on the mutant allele (Figure 6.5F). I found no significant difference in the interprobe distances or co-localisation frequencies between hSHH and hLSC2 between the wild type and mutant alleles (Figure 6.5C, E & G). There was a significant increase in interprobe distance between hSHH and hLSC3-5 on the mutant chromosome in comparison to wild type (Figure 6.5D & E). A reduction in the proportion of co-localised hSHH and hLSC3-5 are also found on the mutant allele (Figure 6.5H). Comparing the wild type and mutant alleles, the differences in interactions are more pronounced than when just comparing the parental and patient cell lines (e.g. in Figure 6.4C ~50 nm change in mean interprobe distance from hSHH to hLSC1 versus ~200 nm change in Figure 6.5B) supporting my initial observations (see 6.2.1). Allele-specific DNA FISH in the heterozygous parental cell lines shows that these disrupted interactions are specific to the chromosome with the 12 kb deletion and supports differences in *SHH* TAD organisation.

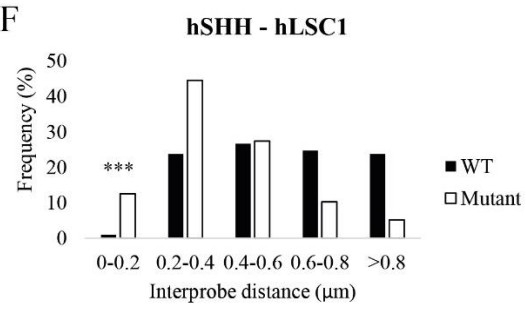
A



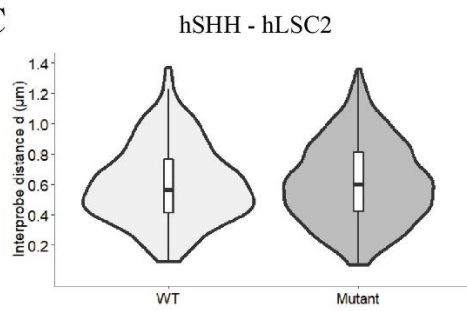
B



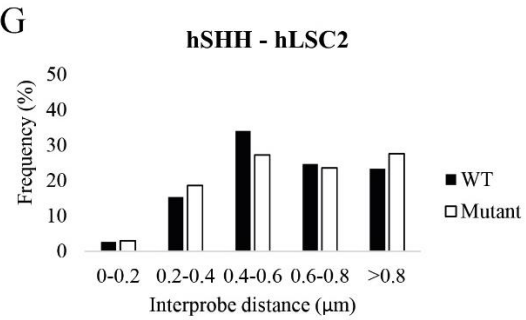
F



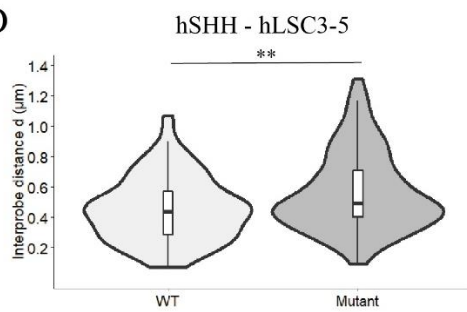
C



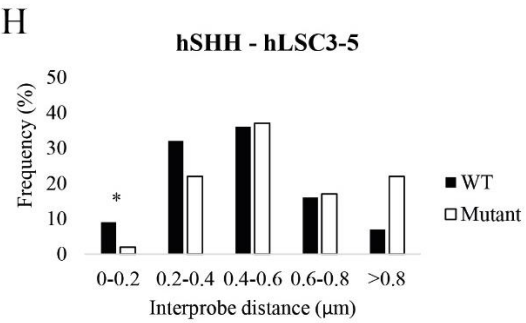
G



D



H



E

Probesets	Wild type vs mutant (p -value)
hSHH - hLSC1	2.08E-11
hSHH - hLSC2	0.7602
hSHH - hLSC3-5	0.004537

Figure 6.5 Altered CTCF looping interactions are specific to the chromosome carrying the 12 kb deletion.

(A) Images of representative nuclei from DNA FISH analysis of parental and patient lymphoblastoid cells showing FISH signals for SHH, LSC2, LSC3-5 and 12kb probes. The 12kb probe is used to distinguish between wild type and mutant alleles. Scale bars: 5 μ m. (B-D) Violin plots show the interprobe distance (μ m) distribution between hSHH-hLSC1 (B), hSHH-hLSC2 (C) and hSHH-hLSC3-5 (D) probes comparing wild type and mutant alleles following DNA FISH of lymphoblastoid cell lines MPR1 (patient), MPR2 (parental) and MPR3 (parental). The statistical significance between data sets was examined by Mann-Whitney U Tests, * < 0.05, ** < 0.01, n = 75-150 alleles. (E) Table summarising the results (p-value) of Mann-Whitney U Tests comparing the interprobe distances between the alleles. (F-H) Frequency distribution of the interprobe distance calculated every 0.2 μ m bin between hSHH-hLSC1 (F), hSHH-hLSC2 (G) and hSHH-hLSC3-5 (H) probes comparing wild type and mutant alleles. The statistical significance between data sets was examined by Fisher's Exact Tests, * < 0.05, *** < 0.001.

6.2.3 SHH TAD structure in primary human fibroblasts

An unexpected observation with the DNA FISH measurements from across the *SHH* TAD was, even on the wild type chromosome, the mean interprobe distance for all sets of fosmid probes measured were between ~450-550 nm (Figure 6.4 & Figure 6.5). Considering the strong peak interaction detected by Hi-C between *SHH*/LSC3-5 within the TAD structure and *SHH*/LSC1 at the sub-TAD border (Figure 6.1B), suggesting frequent interactions, I expected shortened intrachromosomal distances. In comparison, interprobe distances from the *Shh* gene to the ZRS limb enhancer (which show a strong peak interaction by 5C, see 8.2) in either mouse ESCs or tissue are usually within the 200-300 nm range. The shortest median spatial distance detected was from *SHH* to LC3-5 in comparison to LSC1/2 despite these loci having the largest linear distance. This suggests that the *SHH* TAD chromatin is structured in a way that brings these two loci closer together in cells. The large DNA FISH measurements may just be specific to the lymphoblastoid cells used here and the spatial distances may differ in other cell types e.g. human ESCs.

I decided to look at the *SHH* TAD by DNA FISH in another human cell line in order to see if the larger mean interprobe distances were specific to lymphoblastoid or human

cells in general. I obtained fixed slides of primary human fibroblasts from Patricia Heyn (Andrew Jackson Group). I first checked the distribution of CTCF binding across the *SHH* TAD in fibroblasts cells using an ENCODE ChIP-seq track from IMR90 cells (Figure 6.6A). Overall the CTCF distribution is similar between fibroblast and lymphoblastoid cells. There is a CTCF peak detected in IMR90 cells at LSC2 that is absent in lymphoblastoid cells (Figure 6.6A). It should be noted that the IMR90 cells are a fibroblast cell line derived from lung tissue of a 16 week old foetus whereas the primary fibroblasts cells I obtained which are derived from adult skin tissue. I performed 4-colour DNA FISH using probes that hybridised to the *SHH* promoter, LSC2 and LSC3-5.

The range of interprobe distances detected between hSHH and either LSC2 or LSC3-5 in fibroblasts are notably shorter (median distance ~270-400 nm) in comparison to those measured in MPR cell lines (Figure 6.4 & Figure 6.6). This suggests that the larger interprobe distances I found in lymphoblastoid cells may not be a general feature of human cells and could just be specific to the MPR cells used here. I compared the measurements from the wild type allele of lymphoblastoid cells to those from fibroblasts cells. The interprobe distance from hSHH to hLSC2 was significantly shorter in fibroblasts whereas the distances from hSHH to hLSC3-5 were comparable (Figure 6.6B & C). As the probes for the *SHH* gene and LSC2 were found close together in fibroblasts cells, an interaction not seen in the lymphoblastoid cells, I also looked at a published Hi-C heatmap of the IMR90 cell line using data from Rao et al., 2014 as a comparison (Figure 6.6D). Interestingly, a peak interaction is observed between *SHH* and LSC2 as well as with LSC3-5 in fibroblasts. This indicates that CTCF binding at LSC2 facilitates chromatin looping with the *SHH* gene in these cells as well as the *SHH*/LSC3-5 loop found in both cell types. The *SHH*/LSC2 interaction is equivalent to the *Shh*/CTCF3 loop we observe in mouse cells/tissue.

red), embryonic stem cell (H1-hESC, green) and foetal lung fibroblast (IMR90, black) line. The *LMBR1-SHH* CTCF (LSC) sites are numbered below (1-5). **(B)** Violin plots show the interprobe distance (μm) distribution between hSHH-hLSC1, hSHH-hLSC2 and hSHH-hLSC3-5 probes following DNA FISH of primary fibroblast cell line. The statistical significance between the wild type lymphoblastoid allele and fibroblast data sets was examined by Mann-Whitney U Tests, **** < 0.0001 . $n = 75\text{-}150$ alleles. **(C)** Table summarising the results (p -value) of Mann-Whitney U Tests comparing the interprobe distances between the wild type lymphoblastoid allele and fibroblast data sets. **(D)** Hi-C heatmap of the *SHH* TAD from wild type IMR90 foetal lung fibroblast cells at 8-kb resolution (data shown here is from Rao et al., 2014 and was created using HiGlass). Genes (grey) and *LMBR1-SHH* CTCF (LSC) sites 2-5 are numbered above (black) track are shown above.

6.3 Discussion & conclusions

In this chapter, I have presented analysis of the human *SHH* TAD structure in human cell lines that were derived from a patient with acheiropodia, her unaffected parents and from normal adult skin. I showed that the general features of the *SHH* TAD such as the boundary positions and CTCF-mediated chromatin loops are similar to those we detect in mouse but the distribution of CTCF sites and loop anchors differ between the species. I next demonstrated using DNA FISH that loss of a 12 kb region containing three CTCF sites in lymphoblastoid cells from an acheiropodia patient alters the chromatin organisation of the *SHH* TAD. I then compared DNA FISH measurements between wild type lymphoblastoid and primary fibroblast cells and concluded that there were differential interactions within the *SHH* TAD between these two cell types.

6.3.1 Removal of CTCF sites causes acheiropodia in humans

The acheiropodia causative locus was shown to be located to a 12 kb region found in intron 4 of *LMBR1* that contained three sites that are consistently bound by CTCF across many different human cells. These CTCF sites (LSC3-5) appear to function as one of the loop anchors that likely constrain the extent of cohesin-mediated loop extrusion. Loss of this loop anchor would be predicted to perturb interactions between this region and the *SHH* gene. Indeed, I showed that there was an increased interprobe distance between *SHH* and LSC3-5 on the mutant chromosome (Figure 6.5D) and a loss of interaction was also detected by 4C in the MPR1 mutant cell line (Data not shown). We and others have similarly shown in mouse that knockout of CTCF binding sites that mediate the chromatin loop of the *Shh* TAD (Δ CTCF1/2/3, Δ CTCF i4:i5) also perturbs interactions between regions found at the base of the loop (Figure 6.2 & Figure 6.3) (Paliou et al., 2019; Williamson et al., 2019). However, single or double knockout of CTCF sites is insufficient to cause limb defects in mouse unless crossed onto a sensitised background (Paliou et al., 2019; Williamson et al., 2019). Humans are more sensitive to disruptions to SHH levels e.g. heterozygous mutation of *SHH* segregate with HPE patients whereas in mice, loss of both copies of *Shh* recapitulate a severe HPE-like phenotype (Chiang et al., 1996; Roessler et al., 1996). Therefore different requirements for gene dosage between species maybe account for why a

phenotype is only observed in humans and not in the mouse. CTCF binding may be more critical in humans for modulating enhancer-promoter interactions.

6.3.2 Δ 12 kb mutation specifically affects ZRS long-range regulation

The 12 kb mutation specifically disrupts ZRS-mediated limb structures and does not affect the development of any other SHH-regulated tissues. With regards to neural tissue, this is unsurprising as the majority of the SHH enhancers involved in central nervous system development are either proximal to *SHH* or are located within the *SHH* sub-TAD region distinct from the ZRS location. Chromosomes carrying the acheiropodia mutation showed increased interaction with LSC1, the sub-TAD loop anchor, by DNA FISH (Figure 6.5B) and 4C (data not shown). This increased interaction does not seem to have an effect on *SHH* regulation via the neural enhancers located here. A number of long-range enhancers regulating *SHH* expression in epithelial tissues from the oral cavity to the hindgut are located between LSC1 and LSC3-5 (Figure 1.8). Again, no defects in any of these tissues have been found in acheiropodia patients. This suggests that these enhancers are functioning properly and only ZRS is affected by the 12 kb deletion. These enhancers may use other alternative CTCF sites located nearby to mediate their interactions with the *SHH* promoter or may not require direct contact for their function as shown by other studies (Benabdallah et al., 2019; Williamson et al., 2019).

It is still not clear how loss of the CTCF sites in the Δ 12kb deletion specifically prevents ZRS from regulating *SHH* expression in the limb. One potential mechanism would be that the altered CTCF interactions at LSC1 stop the ZRS from directly contacting *SHH* and inducing expression in the limb bud. I have shown in the previous results chapter that in the absence of cohesin, *Shh* expression cannot be induced from its long range enhancers, including ZRS, using TALE-VP64 proteins. Increased CTCF binding at LSC1 could block cohesin extrusion and prevent ZRS from ever (or at too low a frequency) being brought into spatial proximity with the *SHH* gene. However, this still does not explain why the other long-range enhancers found after LSC1 (e.g. SBE5) still work. Alternatively, the absence of LSC3-5 sites means the *SHH*-ZRS is not stabilised and insufficient to induce enough *SHH* expression for normal limb

development. These mechanisms are difficult to test as the acheiropodia mutation cannot be modelled in mice. I conclude from this data that these three CTCF sites are important for ZRS regulation during limb development and other long-range enhancers of *SHH* do not require these sites for their function.

6.3.3 Using differentiated cell lines to model human disease

In this chapter, I used both patient and parental-derived lymphoblastoid cell lines to try to understand the consequences of the 12 kb deletion on *SHH* TAD structure and relate this to the acheiropodia disease mechanism. These cell lines are generated by immortalising B-lymphocytes found in blood samples with Epstein-Barr virus. This is a common approach to make patient-derived cell lines due to the ease of obtaining and accessibility of blood samples. However, acheiropodia is a developmental disorder and *SHH* regulation is likely disrupted in the first trimester of pregnancy during limb development of the foetus and adult B cells are not a relevant cell type for studying *SHH* regulation in the limb but may provide a means to examine differences in *SHH* TAD structure. TAD structure is generally viewed as being conserved across cell types (Dixon et al., 2012; Nora et al., 2012) but how true this statement is has recently been questioned (Eres and Gilad, 2020). In the future, it would be useful to perform DNA FISH on human embryonic tissue or stem cells as a comparison. It is difficult to get human tissue at the correct developmental stage for *SHH* regulation as the majority of available human foetal tissue is from the second trimester and digits have already formed by this point. Comparing wild type and 12 kb mutated human embryonic stem cells may be a more feasible option. An alternative approach to model acheiropodia could involve generating a mouse strain where the entire *SHH* TAD has been humanised. It is now possible to humanise several Mb of DNA in mice although there are still some technical challenges (Zhu et al., 2019).

I also used primary human fibroblasts to look at the structure of the *SHH* TAD. I found that in fibroblasts there were decreased interprobe distances between *SHH* and LSC2 compared to lymphoblastoid cells. This indicates that in fibroblasts a chromatin loop may also form from *SHH* to LSC2 as well as LSC3-5. This cell type difference in CTCF interactions indicates that TAD structure may not necessarily be conserved

between cell types of the same species (Eres and Gilad, 2020). CTCF/RAD21 ChIP-seq carried out by Nadav Ahituv's group on wild type and MPR1 lymphoblastoid cell lines found that in the mutant cells there was ectopic CTCF and cohesin binding at LSC2. However, I did not see altered interactions with LSC2 by DNA FISH despite this site being capable of forming a chromatin loop as observed in the fibroblast cells. Why do some CTCF sites interact with other sites in certain cell types? Again, the lymphoblastoid cells are not the correct cell type here and potentially these sites do form a chromatin loop in ZRS-regulated tissue that may have impair the SHH-ZRS interaction leading to acheiropodia.

6.3.4 Comparison of the mouse and human *SHH* TAD

The *SHH* TAD boundary appears to be conserved between mouse and human. Conversely, many more CTCF peaks are detected by ChIP-seq at the *LMBR1* TAD boundary that could act as boundaries with different strengths. Depending on the algorithms used, different thresholding, window size used etc. mean that where the boundary is called can change (Wit, 2020). One key difference between mouse and human *SHH* TADs is the position of one of the loop anchors of the major internal chromatin loop detected within *LMBR1* of the (CTCF3/i4 in intron 5 of mouse vs LSC3-5 in intron 4 of human) (Figure 6.2). Loss of CTCF3 in mouse perturbs the loop but not *Shh* expression whereas the human 12 kb deletion removes LSC3-5, disrupts the interaction, is associated with acheiropodia and presumably leads to loss of *SHH* expression during limb development. CTCF loops in humans may be more important for mediating cell-type specific regulatory interactions. The abundance of CTCF peaks detected by ChIP-seq in human cells lines makes the data far more difficult to interpret. The species differences in chromatin interactions, and the lack of an acheiropodia-like phenotype in $\Delta 12$ kb mice, highlight the limitations of using mouse models to understand the disease mechanism of some human Mendelian disorders.

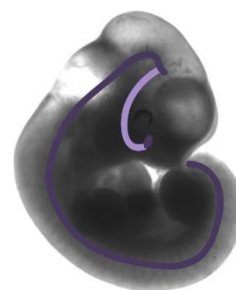
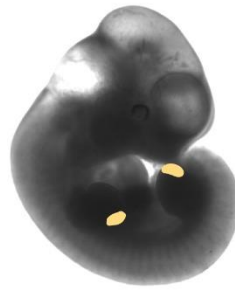
Chapter 7. Concluding remarks

During my PhD, I showed using DNA FISH that in the developing forebrain the forebrain-specific *Shh* enhancer SBE2 specifically co-localises with *Shh* in SBE2-regulated cells (Figure 7.1). This interaction is indicative of a chromatin looping mechanism of enhancer-promoter communication for this enhancer in line with previous observations with the ZRS limb enhancer (Williamson et al., 2016) but in contrast to another *Shh* neural enhancer SBE6 (Benabdallah et al., 2019). The disparity between the modes of enhancer-promoter communication used by these enhancers may be a reflection of distance to the *Shh* gene. Evidence from 3C studies show that the more distally located epithelial enhancers directly interact with the *Shh* promoter in a tissue-specific manner (Seo et al., 2018; Tsukiji et al., 2014). In the future it would be interesting to test using DNA FISH if the SLGE enhancer (100 kb upstream of *Shh* similar to SBE6) shows an increased distance from the *Shh* promoter like SBE6 in the tissue where this enhancer regulates *Shh* expression.

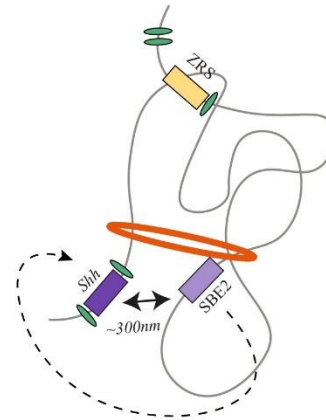
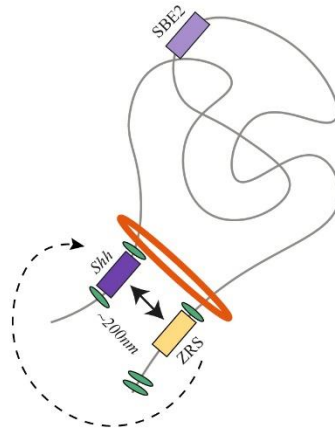
I investigated the role of CTCF binding in chromatin structure and distal regulation of *Shh*. In Δ CTCF mouse models, I found using DNA FISH that loss of single CTCF binding sites disrupted the chromatin structure between the *Shh* gene and its enhancer in a tissue-specific manner leading to increased separation between the loci (Figure 7.1). Perturbation to *Shh* TAD structure and enhancer-promoter spatial proximity by loss of single CTCF binding sites was insufficient to disrupt developmental *Shh* regulation. Using a synthetic enhancer system in mESCs, I showed that acute loss of CTCF binding genome-wide did not prevent *Shh* induction, assayed by nascent RNA FISH, following enhancer activation with TALE-VP64 proteins. These findings support the role of CTCF in the structuring of chromatin into TADs as we showed by 5C that TAD boundaries, intra- and inter-TAD interactions were altered in Δ CTCF mESCs (Williamson et al., 2019). Here, I found that *Shh* expression was not dependent on CTCF binding. CTCF binding may function as an additional regulatory layer that exists to maximise transcriptional output by increasing the likelihood enhancers and promoter will interact within a spatial domain or TAD but is not directly involved in gene activation *per se*. Further genetic manipulation of the *Shh* TAD to remove all CTCF binding sites and assess developmental gene regulation would be an interesting future avenue for this project.

**ZRS-regulated
limb tissue**

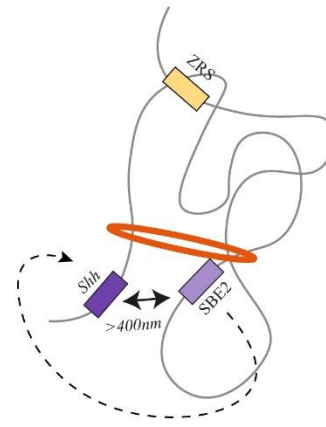
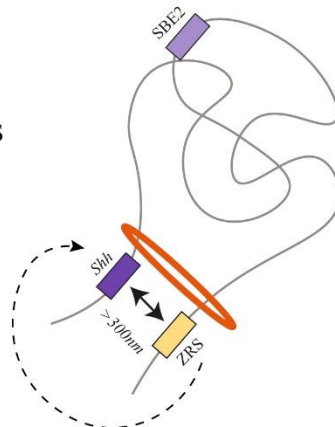
**SBE2-regulated
forebrain tissue**



1. Tissue-specific interactions occur between distal enhancers and *Shh* during embryogenesis



2. Chromatin structure is altered by CTCF loss but *Shh* expression is unperturbed



3. Cohesin is required for *Shh* enhancer-driven gene activation

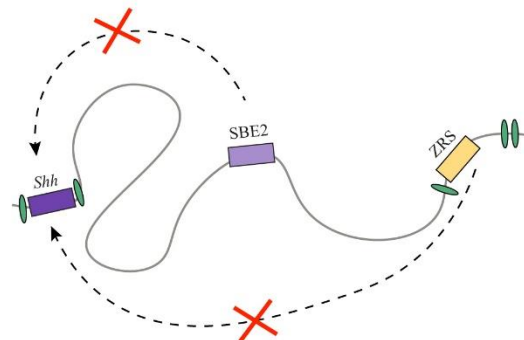


Figure 7.1 The relationship between chromatin structure and distal regulation.

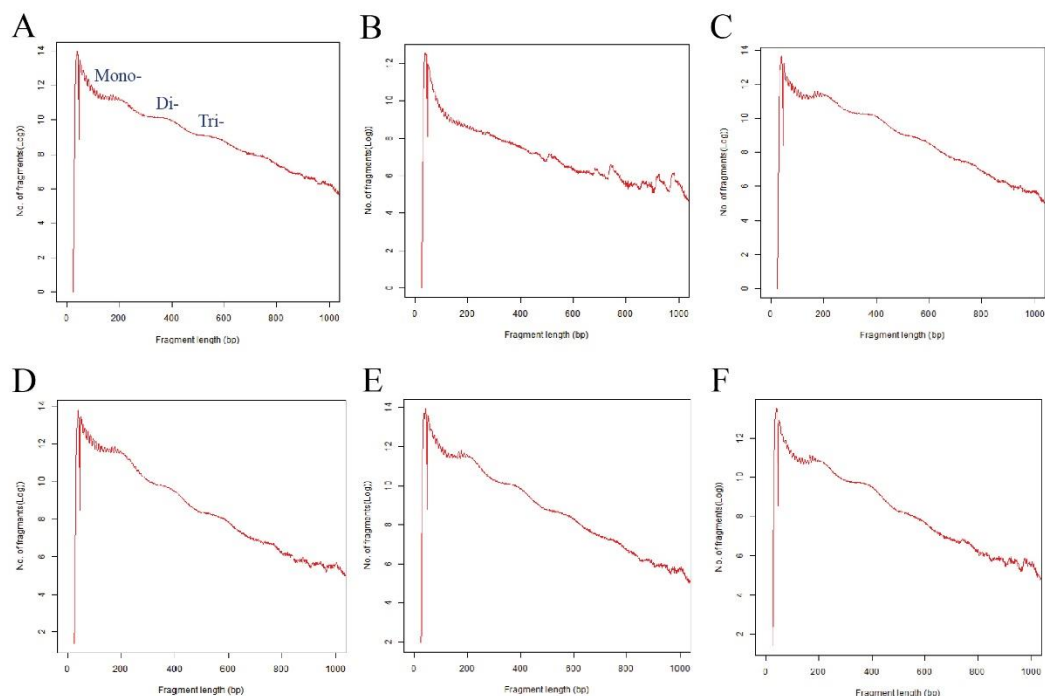
Summary of the key findings of this thesis. (1) Tissue-specific co-localisation between *Shh* and the forebrain enhancer SBE2 was detected in SBE2-regulated forebrain tissue similar to previous observations in ZRS-regulated limb tissue. (2) Single CTCF site knockouts disrupted *Shh* TAD structure, reduced tissue-specific enhancer-promoter co-localisation but did not affect developmental *Shh* expression. *Shh* expression could still be induced by targeting distal enhancers with synthetic activators in mESCs following acute depletion of CTCF genome-wide. (3) Following genome-wide depletion of the cohesin complex by acute degradation of SCC1, *Shh* expression could not be induced using synthetic activators targeting *Shh* distal enhancers in mESCs.

I used acute depletion of SCC1 in combination with synthetic transcription factors to study the role of the cohesin complex in distal activation of *Shh*. In the absence of cohesin, *Shh* nascent expression was markedly reduced from its distal enhancers (Figure 7.1). This finding supports that cohesin-mediated loop extrusion allows enhancers located hundreds of kilobases away in the linear genome to activate their target gene by coming into close spatial proximity. I propose that the presence of cohesin-mediated extrusion explains why in Δ CTCF tissue or CTCF-depleted cells *Shh* expression was still observed as the presence of cohesin still facilitated enhancer-promoter interactions albeit these will now be transient interactions. It is unclear whether all *Shh* distal enhancers rely on cohesin for enhancer-driven gene activation. As SBE6 shows increased spatial separation from the *Shh* promoter when its activated, suggestive of a distinct model of enhancer-promoter communication in contrast to chromatin looping indicated by SBE2/ZRS enhancers, in the future I will test if SBE6 also depends on cohesin for *Shh* induction.

Altogether, these findings suggest a revised outlook on the role of TADs in distal regulation. Cohesin-mediated loop extrusion is required to form favourable interactions between long-range enhancers and promoters that leads to gene activation and CTCF functions to limits this extrusion process and therefore the spatial search space of regulatory elements within a TAD.

Chapter 8. Appendices

8.1 Additional Figures

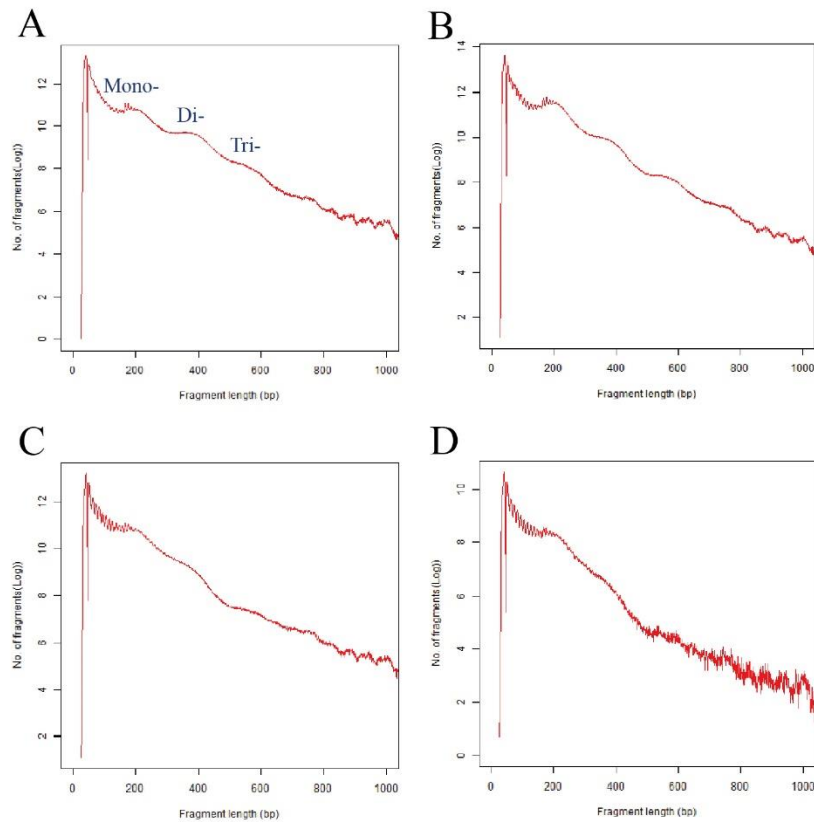


G

E11.5 ShhGFP X B16	GFP+			GFP-		
Replicate	1	2	3	1	2	3
Read-pairs before clipping:	94100917	58226055	63590926	68794681	72833426	55219488
Read-pairs after clipping:	89455488	54379607	60655805	65757757	69533359	52889688
all read-pairs aligned:	82593632	29533848	50617794	61892078	62288028	42839447
Mitochondrial read-pairs in aligned:	3200368	382002	0	3779053	0	0
Nuclear read-pairs after deduplication:	45263093	7642599	36786475	43878288	48855357	27463823
% chrM	3.874836	1.293438	0	6.105875	0	0
% aligned	92.32931	54.31052	83.45087	94.12133	89.58006	80.99773
% nuclear	48.10059	13.12574	57.84862	63.78151	67.07821	49.73574

Appendix Figure 1. QC analysis of E11.5 ATAC samples following sequencing.

(A-F) Fragment length distribution profiles following sequencing of libraries prepared from E11.5 isolated neural cells from GFP+ replicate (A) 1, (B) 2, and (C) 3 and GFP- replicate (D) 1, (E) 2 and (F) 3. Peaks are shown for mono-, di- and tri-nucleosomal fragments (bp). (G) Summary of the read-pairs obtained for each ATAC-seq sample including the number of read-pairs before and after clipping and the number and percentage of mitochondrial, aligned and nuclear read-pairs after removal of duplicates.



E

E14.5 ShhGFP X Bl6	GFP+		GFP-	
Replicate	1	2	1	2
Read-pairs before clipping:	57374221	63465878	5.05E+07	35608664
Read-pairs after clipping:	54771134	59273934	48608532	34260377
all read-pairs aligned:	45247244	56038576	38521923	2073993
Mitochondrial read-pairs in aligned:	0	5451449	0	0
Nuclear read-pairs after deduplication:	25960334	38744308	21939907	1749327
% chrM	0	9.728029	0	0
% aligned	82.61148	94.54169	79.2493	6.05362
% nuclear	45.24738	61.04746	43.42404	4.912644

Appendix Figure 2. QC analysis of E14.5 ATAC samples following sequencing.

(A-F) Fragment length distribution profiles following sequencing of libraries prepared from E14.5 isolated neural cells from GFP+ replicate (A) 1, (B) 2, and (C) 3 and GFP- replicate (D) 1, (E) 2 and (F) 3. Peaks are shown for mono-, di- and tri-nucleosomal fragments (bp). (G) Summary of the read-pairs obtained for each ATAC-seq sample including the number of read-pairs before and after clipping and the number and percentage of mitochondrial, aligned and nuclear read-pairs after removal of duplicates.

Rank	<i>de novo</i> motif (forward)	<i>P</i> -value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-2567	CTCF(L) YY1 NeuroD1 Zic1/4 Sp4 Atoh1	3785 (8.51%)	330 (0.75%)
2		1.00E-295	Sp1/2/4 Klf4/5/7/14	9568 (21.51%)	6613.8 (14.96%)
3		1.00E-253	NfA/B NfC/X Dux Mx1/2 Tn2	3588 (8.06%)	1940.8 (4.39%)
4		1.00E-219	Rfx1/2/3/4/5	1169 (2.63%)	394 (0.89%)
5		1.00E-217	Rarg/Rar/Rxr Crx Foxk1 Mybl1 Gsc Dmbx1	165 (0.37%)	3.3 (0.01%)
6		1.00E-190	Mtf11 Irfl/4/7 Foxa2 Prdm1 Rbpj1	116 (0.26%)	1.4 (0%)
7		1.00E-188	Nfatc2/3/5 Irx3/4 Six1 Tcfap2e Amt	115 (0.26%)	0.8 (0%)
8		1.00E-184	Irfl/2/3/7 Stat1/2 Myb Six1	113 (0.25%)	1.9 (0%)
9		1.00E-119	E2f1/2/3 Runx2/3 Mafk Klf7	19707 (44.3%)	17185.4 (38.87%)
10		1.00E-106	YY1/2 Arid5a Neurod2 Neurog1/2 Atoh1 Bhlha15	97 (0.22%)	3.9 (0.01%)
13		1.00E-95	Sox2/3/4/9/10/11	4663 (10.48%)	3414.9 (7.72%)
14		1.00E-90	Elk1/3/4 Fhl1 Ets1/4 Erg Fev	11028 (24.79%)	9197.4 (20.8%)

Rank	<i>de novo</i> motif (forward)	<i>P</i> -value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-2007	CTCF(L) NeuroD1 Zic1/3/4 Sp4	3949 (10.03%)	534.8 (1.37%)
2		1.00E-343	Sp1/2/4 Klf4/5/7/14	13492 (34.28%)	9889.2 (25.29%)
3		1.00E-294	NfA/B Dux Mx1/2 NfC/X	3763 (9.56%)	1969.5 (5.04%)
4		1.00E-288	Bbx Mybl1 Sox1 Crx Otx2 Pitx2	219 (0.56%)	4.7 (0.01%)
5		1.00E-179	Rfx1/2/3/4/5/7	1594 (4.05%)	710.6 (1.82%)
6		1.00E-177	Mtf11 Irfl/7 Hbp1 Prdm1 Sp1 Foxa2 Pou6f1 Spic	128 (0.33%)	2 (0.01%)
7		1.00E-162	Irfl/2/3/7 Stat1/2 Myb Six1	102 (0.26%)	0 (0%)
8		1.00E-147	YY2 Atf1 Neurog1/2 Neurod2 Gfx Arid5a Atoh1 Cdx1	111 (0.28%)	3 (0.01%)
9		1.00E-145	Nfatc1/2/3/5 Mathk Ehf Chr	110 (0.28%)	2.1 (0.01%)
10		1.00E-128	Arid5a Ddit3 Tead3/4 Tcf7 Nfatc2	111 (0.28%)	3.5 (0.01%)
11		1.00E-123	E2f2/3 Obox6 Otx1 Dmbx1 Pou6f1/2 YY2 GSC2	16998 (43.18%)	14605.3 (37.35%)
14		1.00E-84	Elk1/4 Sp1 Nfatc1/3 Fhl1 Efl1 Erg	9364 (23.79%)	7728.3 (19.76%)

Appendix Figure 3. Motif analysis of E10.5 unique peaks

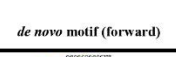
Summary of top 10 enriched motifs and any other motifs that were found in >5% of peaks identified from HOMER *de novo* motif output from E10.5 *Shh*-expressing (top) and non-expressing (bottom) cells showing the associated transcription factor family, a representative profile of the binding motif and the cell populations associated with a particular motif.

Rank	<i>de novo</i> motif (forward)	P-value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-2710	CTCF(L) NeuroD1 Zic1/3/4 Sp4 Zn354C	5478 (11.5%)	772 (1.63%)
2		1.00E-425	YY1/2 BREu NeuroD1 Atf3 Gata3 E2f3 Xcep1 Tcf2a	1434 (3.01%)	339.9 (0.72%)
3		1.00E-402	Id4 Myf6 Snai2 Usf2 E2a Hey1/2 Myod1 Pknx2 Hes5	1120 (2.35%)	221 (0.47%)
4		1.00E-399	Sp1/2/4 Klf4/5/7/14	11670 (24.49%)	7919.5 (16.77%)
5		1.00E-347	NfyA/B Dux(B) Bapx1 Nkx2.2/5 Nfic	9112 (19.12%)	5979.7 (12.66%)
6		1.00E-344	Smad3 E2f1/4/6/7/8 Brex Zbtb7b	927 (1.95%)	176.4 (0.37%)
7		1.00E-316	Elk1/4 Bcl6b Zfp410 Hand1 Etv1/4 Smad3 Gabpa	1061 (2.23%)	249.5 (0.53%)
8		1.00E-304	Hic1 Hand1 Zbtb18 Rorgt E2f1/2/3 Smad2/4 Tcf2a	1143 (2.4%)	296.9 (0.63%)
9		1.00E-295	E2a Smi2 Zbtb7A/B YY1 Id4 Tbx5 Tcf3 Figla	1514 (3.18%)	495.8 (1.05%)
10		1.00E-288	Brex Zbtb7A/C Plagl1 E2f1/2/3 YY2 Zn165 Rara	773 (1.62%)	146.4 (0.31%)
11		1.00E-231	Elk1/3 E1f1 Etv1/4 Fev Gabpa	4743 (9.95%)	2873.6 (6.09%)

Rank	<i>de novo</i> motif (forward)	P-value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-5519	CTCF(L) NeuroD1 Znf923 Zic1/4 Zn354C Sp4	6984 (10.86%)	463.8 (0.73%)
2		1.00E-340	Sp1/2/4 Klf7/12/14 Bcl6b Zfp410	21193 (32.95%)	16507.1 (25.95%)
3		1.00E-305	En1/2 Lhx8 Mx1 Prrx2 Lbx2 Rax2 NfyB Gbx1	10879 (16.92%)	7560.9 (11.89%)
4		1.00E-296	Rfx1/2/3/4/5	1605 (2.5%)	541.7 (0.85%)
5		1.00E-212	Etv6 CebpB/D/E/G Elk1/4 E1f1 Atf1 Ehf	26236 (40.80%)	22184.3 (34.87%)
6		1.00E-197	Crx Pitx1/3 Dmbx1 Gsc Otx2	175 (0.27%)	5.7 (0.01%)
7		1.00E-193	Runx2/3 Bmyb Otx1 Obox6 Tfc2p Dmbx1 Pou6f2 Grhl1	26185 (40.72%)	22304.9 (35.06%)
8		1.00E-186	Tcfap2e Hif2a Id2 Hif1b Nfat5 Arnt1 Nkx2.2 Nfatc2	114 (0.18%)	1.2 (0%)
9		1.00E-170	Irfl/2/3/7 Stat1 Ttp Six1	106 (0.16%)	0.9 (0%)
10		1.00E-155	Neurod2 Bhlha15 Atoh1 Twis2 Neurog1/2 YY2 Olig2/3	99 (0.15%)	1.9 (0%)
11		1.00E-116	Zbtb3 Ddit3 Inr MatB/K CTCF Zn143 Gm397 Cebpa	10957 (17.04%)	8785.7 (13.81%)
12		1.00E-114	Sox2/3/4/9/10/11	3975 (6.18%)	2702.7 (4.25%)
13		1.00E-105	Creb1/5 Atf1/2/7 Jundm2 Jdp2	4513 (7.02%)	3198.5 (5.03%)

Appendix Figure 4. Motif analysis of E11.5 unique peaks

Summary of top 10 enriched motifs and any other motifs that were found in >5% of peaks identified from HOMER *de novo* motif output from E11.5 *Shh*-expressing (top) and non-expressing (bottom) cells showing the associated transcription factor family, a representative profile of the binding motif and the cell populations associated with a particular motif.

Rank	<i>de novo</i> motif (forward)	<i>P</i> -value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-6314	CTCF(L) NeuroD1 Zic1/3/4 Sp4 ZNF354C	10154 (14.92%)	1042.6 (1.55%)
2		1.00E-465	Rfx1/2/3/4/5	2480 (3.64%)	828.7 (1.23%)
3		1.00E-364	Sp1/2/4 Klf4/5/7/14/16	15230 (22.38%)	10968.9 (16.33)
4		1.00E-360	Lhx1/2/3/6/8 Vsx1/2 Emx2 Noto	10905 (16.03%)	7304 (10.88%)
5		1.00E-319	Sox2/3/4/6/9/10/11	19200 (28.22%)	14745.7 (21.96%)
6		1.00E-304	Etv1/4/5 Elk1/4 Fev Fli1 Erg Sp1 Ets1	31250 (45.92%)	26112.8 (38.88%)
7		1.00E-205	Zbtb3 Ddit3 Inr Znf143 MatB/K CTCF Pou3f3 Cebpa Sox4	20229 (29.73%)	16495.6 (24.56%)
8		1.00E-200	Nfatc1/2/3/5 Amt Creb3 Hif1B/2A	121 (0.18%)	2 (0%)
9		1.00E-190	Crx Dmbx1 Otx2 Pitx1/3 Gsc2	161 (0.24%)	4.4 (0.01%)
10		1.00E-173	Nkx2.1/2/3/5/8 Bapx1 Nkx3.1/2	108 (0.16%)	1.7 (0%)
20		1.00E-55	Meis1/2/3 Tgfr1/2 Pknx2 Smad2	9866 (14.5%)	8366.9 (12.46%)
23		1.00E-29	Tead1/3/4 Thx4/5 Tbet Klf10 Runx1 Lomes Gm397	6168 (9.06%)	5283.4 (7.87%)

Rank	<i>de novo</i> motif (forward)	<i>P</i> -value	Predicted to be bound by	No. and % of peaks with motif	No. and % of background with motif
1		1.00E-881	Sp1/2/4 Klf4/5/7/14	6396 (36.67%)	5285.3 (16.61%)
2		1.00E-792	NfyA/B Dux NRC/X Mx1/2 Duxbl	3134 (17.97%)	1652.1 (5.19%)
3		1.00E-524	CTCF(L) Zic1/3/4 NeuroD1 Sp4	1502 (8.61%)	574.2 (1.8%)
4		1.00E-399	Elk1/3/4 Efl1 Etv1/4 Fev Erg Fli1	11670 (24.49%)	7919.5 (16.77%)
5		1.00E-266	Gly Runx1/2 Irc Mzf1 Zbtb3 Foxj1	667 (3.82%)	217.3 (0.68%)
6		1.00E-251	Crx Pitx1/3 Gsc2 Otx2 Dmbx1	208 (1.19%)	9.9 (0.03%)
7		1.00E-211	Creb1/5 Atf1/7 Mlx Srebf1/2 Bhlhe41	2535 (14.54%)	2413.1 (7.58%)
8		1.00E-209	Tead1/2/3/4 Nfatc1/2/3 Neurog2	111 (0.64%)	1.2 (0%)
9		1.00E-200	YY1/2 Arid5a Neurog1/2 Neurod2 Atoh1 Olig3	107 (0.61%)	1.3 (0%)
10		1.00E-191	Nrf1 Zfp161 Hes1 Arnt Egr2/3 Tcf15 Znf165	1544 (8.85%)	1224.8 (3.85%)

Appendix Figure 5. Motif analysis of E14.5 unique peaks

Summary of top 10 enriched motifs and any other motifs that were found in >5% of peaks identified from HOMER *de novo* motif output from E14.5 *Shh*-expressing (top) and non-expressing (bottom) cells showing the associated transcription factor family, a representative profile of the binding motif and the cell populations associated with a particular motif.

8.2 Publication

© 2019. Published by The Company of Biologists Ltd | Development (2019) 146, dev179523. doi:10.1242/dev.179523



RESEARCH ARTICLE

Developmentally regulated *Shh* expression is robust to TAD perturbations

Iain Williamson, Lauren Kane, Paul S. Devenney, Ilya M. Flyamer, Eve Anderson*, Fiona Kilanowski, Robert E. Hill, Wendy A. Bickmore[‡] and Laura A. Lettice[‡]

ABSTRACT

Topologically associating domains (TADs) have been proposed to both guide and constrain enhancer activity. *Shh* is located within a TAD known to contain all its enhancers. To investigate the importance of chromatin conformation and TAD integrity on developmental gene regulation, we have manipulated the *Shh* TAD – creating internal deletions, deleting CTCF sites, and deleting and inverting sequences at TAD boundaries. Chromosome conformation capture and fluorescence *in situ* hybridisation assays were used to investigate the changes in chromatin conformation that result from these manipulations. Our data suggest that these substantial alterations in TAD structure have no readily detectable effect on *Shh* expression patterns or levels of *Shh* expression during development – except where enhancers are deleted – and result in no detectable phenotypes. Only in the case of a larger deletion at one TAD boundary could ectopic influence of the *Shh* limb enhancer be detected on a gene (*Mnx1*) in the neighbouring TAD. Our data suggests that, contrary to expectations, the developmental regulation of *Shh* expression is remarkably robust to TAD perturbations.

KEY WORDS: Chromosome conformation capture, Fluorescence *in situ* hybridisation, *Shh*, Enhancers, Genome organisation, Development, Topologically associated domains

INTRODUCTION

At the megabase-scale, the mammalian genome is partitioned into self-interacting topologically associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012). Mammalian TAD boundaries are enriched in CTCF sites, with their relative orientation appearing crucial to function (Narendra et al., 2015; Rao et al., 2014; Sanborn et al., 2015). TADs are formed by dynamic cohesin-driven loop extrusion (Fudenberg et al., 2016; Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Vian et al., 2018) and convergent CTCF sites act to impede loop extrusion by enabling WAPL-mediated release of cohesin from the chromosome (Haarhuis et al., 2017).

The regulatory landscapes of developmental genes are frequently found to be contained within the same TAD (Dixon et al., 2012; Rao

et al., 2014). TADs have, therefore, been proposed to act as functional regulatory units within which contacts between enhancers and their target gene are favoured and aberrant interactions of enhancers across TAD boundaries are limited (Fudenberg et al., 2016; Sun et al., 2019). In support of this hypothesis, some studies have found that deletion or inversion of CTCF sites at TAD boundaries can promote TAD boundary crosstalk and rewire enhancer-promoter contacts (de Wit et al., 2015; Guo et al., 2015; Narendra et al., 2015; Rodríguez-Carballo et al., 2017). Moreover, a number of recent studies have suggested that changes to TAD structure can disrupt gene regulation through enhancer-rewiring in human disease (Flavahan et al., 2016; Franke et al., 2016; Lupiáñez et al., 2015). However, other studies report that, although depletion of CTCF erases the insulation between TADs, it has limited effects on gene expression (Nora et al., 2017; Soshnikova et al., 2010).

To further study the CTCF-mediated function of TADs in developmental gene regulation, we have exploited the sonic hedgehog (*Shh*) regulatory domain – a paradigm locus for long-range regulation. The SHH morphogen controls the growth and patterning of many tissues during embryonic development, including the brain, neural tube and limbs. Spatial and temporal *Shh* expression is regulated by tissue-specific enhancers located within the gene, and upstream in a large gene desert and within neighbouring genes (Jeong et al., 2006; Anderson and Hill, 2014). *Shh* and its *cis*-acting elements are all contained within a well-characterised ~960 kb TAD (Anderson et al., 2014; Williamson et al., 2016). In the developing mouse limb bud, *Shh* expression is solely determined by the ZRS (also known as MFCS1) enhancer (Lettice et al., 2003; Sagai et al., 2005) located 850 kb upstream of *Shh* within an intron of the widely expressed *Lmbr1* (Fig. 1A). Fluorescence *in situ* hybridisation (FISH) showed that *Shh* and the ZRS are consistently located in relatively close proximity to each other in all cell types and tissues examined, which we infer to be a consequence of the underlying invariant TAD structure. In contrast, we observed increased ZRS-*Shh* colocalisation in the *Shh*-expressing posterior portion of developing limb buds (Williamson et al., 2016), consistent with a specific gene-enhancer contact.

Here, we genetically manipulate the *Shh* TAD and its TAD boundaries to investigate the importance of chromatin architecture on TAD structure and on the regulation of gene expression. We use a chromosome conformation assay (5C) and FISH to investigate how these manipulations affect structures within the *Shh* TAD and its interactions with adjacent TADs. We determine how these alterations affect the expression pattern of *Shh* and nearby developmentally regulated genes *in vivo*. We also examine the phenotypic consequences of these manipulations. Our results question the importance of TADs for correct spatial and temporal gene regulation.

MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK.

[‡]Present address: Transgenic Technology, CRUK Beatson Institute, Garscube Estate, Switchback Road, Bearsden, Glasgow, G61 1BD, UK.

[‡]Authors for correspondence (Wendy.Bickmore@igmm.ed.ac.uk; Laura.Lettice@igmm.ed.ac.uk)

L.K., 0000-0002-7210-7160; W.A.B., 0000-0001-6660-7735; L.A.L., 0000-0002-4294-8970

Received 15 April 2019; Accepted 2 September 2019

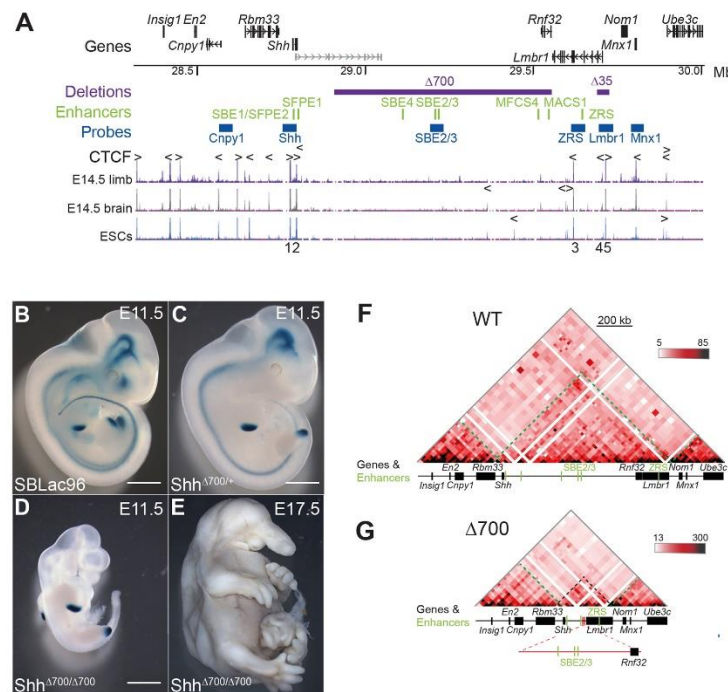


Fig. 1. A 700-kb intra-TAD deletion has no adverse effects on limb-specific expression of *Shh*. (A) The top schematic shows the location of genes in the 1.7 Mb murine genomic locus (chr5: 28317087-30005000; mm9) containing *Shh* analysed by 5C and FISH, with the position of tissue-specific *Shh* enhancers shown below in green. Locations to which the fosmid FISH probes hybridise are shown in blue, and the purple bars indicate deleted genomic regions ($\Delta 700$ and $\Delta 35$). The bottom three tracks show UCSC ENCODE CTCF ChIP-seq profiles in E14.5 limb buds, E14.5 brain and in ESCs. Arrowheads above the tracks indicate the orientation of CTCF-binding motifs, and the deleted CTCF binding sites (1-5) are numbered below. (B-D) β -gal staining to show the expression of the *lacZ* reporter gene carried by the SB transposon in E11.5 embryos. The embryo in B carries the intact SBLac96 insertion (Anderson et al., 2014), whereas C shows the remaining sites of expression after Cre-mediated deletion of 700 kb (purple bar in A) on one allele and D shows the phenotype of an embryo homozygous for the 700 kb deletion, with β -gal staining only evident in the ZPA of the limb buds. Scale bars: 1 mm. (E) Homozygous deletion embryo at E17.5 showing craniofacial and brain defects but normally formed limbs. (F,G) 5C heatmaps from cells of the bodies of E11.5 wild-type embryos (F) and embryos homozygous for the 700 kb deletion (G). Heatmap intensities represent the average of interaction frequency for each window, binned over 21 kb windows, with read counts summed from two biological replicates and colour-coded according to the scale shown. Green dashed lines highlight TAD boundaries, black dashed lines indicate the *Shh* TAD boundaries and the reduced size of the TAD homozygote cells. Data for individual biological replicates are in Fig. S1.

RESULTS

A large deletion within the *Shh* TAD does not disrupt local genome organisation or limb-specific activation of *Shh*

Prominent features of the *Shh* TAD include five CTCF binding sites preserved across multiple cell types (Fig. 1A), and two sub-TADs with overlapping boundaries located within the gene desert between the forebrain enhancers and *Rnf32* (Figs 1F and 2A). This region of the gene desert includes less well defined CTCF peaks that differ across cell types but because of their location may have some role in defining these sub-TADs (Fig. 1A) (Rosenbloom et al., 2013).

To determine the contribution of TAD internal sequence to 3D chromatin organisation and gene expression, we exploited our previous work that used the local hopping activity of the sleeping beauty (SB) transposon to probe the *Shh* regulatory domain (Anderson et al., 2014). Transposition of the SB leaves one *loxP* site at the initial integration site and inserts a second site where it

re-integrates, enabling Cre recombinase to create deletions of the intervening DNA. The orientation of the re-integration means the *lacZ* gene [encoding β -galactosidase (β -gal)] carried by the SB is retained in the deleted chromosome allowing remaining enhancer activity to be monitored. Using this approach, we deleted ~ 700 kb ($\sim 70\%$) of the internal *Shh* TAD sequence, including the sub-TAD boundaries but leaving the five CTCF binding sites at the TAD extremities intact (Fig. 1A). The $\Delta 700$ deletion removes many of the known *Shh* enhancers and relocates the ZRS to within 96 kb of the *Shh* promoter (Fig. 1A). Removal of the *Shh* forebrain and epithelial enhancers in the $\Delta 700$ deletion is shown by changes in the β -gal staining of *Shh* ^{$\Delta 700/+$} embryos. Staining is observed only within the floor plate and hind brain, presumably driven by the proximal enhancers SFPE1/2 and SBE1, and within the limbs driven by the ZRS (compare the wild type in Fig. 1B with the *Shh* ^{$\Delta 700/+$} embryo in Fig. 1C). Homozygous *Shh* ^{$\Delta 700/\Delta 700$} embryos show phenotypes

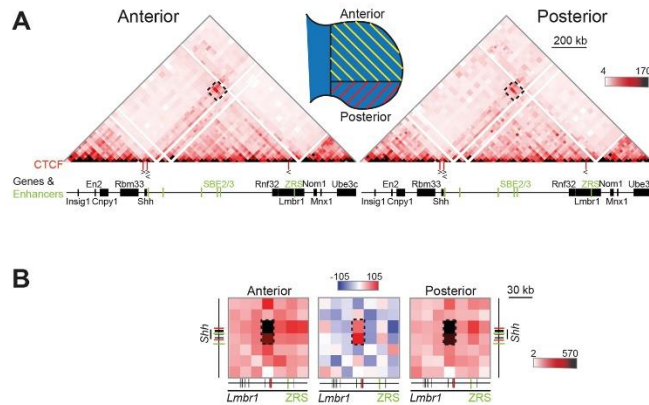


Fig. 2. 5C analysis in E11.5 distal anterior and posterior limb tissue. (A) 5C heatmaps showing data from distal anterior and posterior limb bud cells of E11.5 embryos, across the *Shh* region shown in Fig. 1A. The schematic indicates the limb bud portions dissected for anterior and posterior cell populations. Interactions highlighted by the black dashed boxes locate the region of the heatmaps shown at higher resolution in B. (B) Higher resolution (15 kb binning) 5C heatmaps showing interactions encompassing *Shh* and ZRS. The comparison heatmap (centre), with compared datasets normalised by read count, shows interactions enriched in posterior cells (red) or anterior cells (blue). Enriched interactions between loci containing CTCF binding sites are indicated by the black dashed boxes. Data for individual biological replicates are in Fig. S2.

very similar to those of *Shh*^{-/-} embryos but with normal limb and digit patterning (Chiang et al., 1996). These data indicate that, despite its incorrect position now only 96 kb from the *Shh* promoter, ZRS is able to function normally to drive *Shh* expression in limb development (Fig. 1D,E).

5C on whole embryonic day (E) 11.5 *Shh*^{Δ700/Δ700} and wild-type embryos shows that the *Shh* TAD boundaries and the adjacent TADs are unaffected by the Δ700 deletion (Fig. 1F,G; Fig. S1). Therefore, neither sequence elements nor chromatin interactions within the deleted region is needed for maintaining the location of the TAD boundaries. In addition, the large genomic distance between *Shh* and its limb enhancer ZRS is not required for correct function.

Interactions within the *Shh* TAD are delineated by CTCF sites either side of *Shh* and within *Lmbr1*

Our previous 5C analyses on cells dissected from whole limbs, bodies and heads of E11.5 embryos showed enriched interactions between the genomic region containing *Shh*, located at one TAD boundary, and a genomic region within *Lmbr1* close to ZRS, located ~70 kb from the other TAD boundary (Williamson et al., 2016). That this enrichment can be identified throughout the E11.5 embryo, a stage when we have shown that high levels of *Shh*-ZRS colocalisation occur only in the posterior distal limb, excludes active *Shh*-ZRS colocalisation as the sole driver of this apparent chromatin loop (Williamson et al., 2016). To gain further insight into the nature of these interactions, we dissected E11.5 limb buds to compare cell populations with no ZRS activity (anterior two-thirds of bud) with those in which ZRS is active (posterior one-third) (Fig. 2A; Fig. S2A,C).

The 5C heatmaps of the *Shh* TAD are similar in both anterior and posterior limb bud cell populations, and comparable with dissected E11.5 bodies (compare Fig. 1F with Fig. 2A). At high (15 kb) resolution, the strongest enrichment involved interactions between both *Shh* and the genomic region immediately 3' of *Shh*, and a locus ~20 kb from ZRS in intron 5 of *Lmbr1* (Fig. 2B; Fig. S2B,D, left- and right-hand heatmaps). ENCODE data (Rosenbloom et al., 2013) indicates these three loci are all bound by CTCF across a range of cell and tissue types (Fig. 1A), with the underlying DNA containing CTCF-binding motifs in a convergent orientation consistent with a role blocking loop extrusion (Fig. 2A).

CTCF site deletions reduce *Shh* intra-TAD interactions and disrupts *Shh*/ZRS proximity

To examine the role of CTCF sites on the architecture of the *Shh* TAD we used CRISPR-Cas9 to make small (~1 kb) deletions of sequences containing the five major CTCF binding sites in mouse embryonic stem cells (ESCs) (Table S1). We first generated ESC lines homozygous for deletions of the CTCF binding regions 3' and 5' of *Shh* (Fig. 1A, sites 1 and 2, respectively) and assayed chromatin conformation by 5C and FISH.

The *Shh* TAD structure in wild-type ESCs is similar to that in E11.5 embryos (Fig. 3A; Fig. S3A,C). Deletion of CTCF site 1 (ΔCTCF1), which delineates the TAD boundary 3' of *Shh*, results in *Shh* losing interactions (arrows) with the rest of its own TAD and gaining interactions (arrowheads) with regions just 5' of *En2* and *Rbm33* (Fig. 3A; Fig. S3A,C). The TAD boundary re-locates by ~60 kb to 5' of *Shh* beyond CTCF2 (Fig. 3B,C; Fig. S3B,D). There is also loss of interactions with a locus upstream of the forebrain enhancers near the sub-TAD boundary within the larger *Shh* sub-TAD (Fig. 3B, strong blue diagonals located between SBE3 and *Rnf32*). These data are consistent with CTCF1 forming the *Shh* TAD boundary by blocking loop extrusion emanating from within the *En2* TAD.

The left hand *Shh* TAD boundary is not affected by deletion of CTCF site 2; however, the 5' *Shh* region does gain contacts (arrowheads) with the *En2* TAD in a similar manner to the loss of CTCF1 (Fig. 3A-C; Fig. S3A-D), suggesting that both CTCF1 and 2 are necessary to optimally block loop extrusion emanating from the *En2* TAD. Although loss of CTCF2 results in decreased interactions between the *Shh* locus and the rest of its own TAD (arrows), this is compensated for by increased interactions with CTCF1, thereby maintaining the boundary position (Fig. 3A,B; Fig. S3, right-hand heatmaps). There are also enriched interactions within the *Shh* sub-TAD in ΔCTCF2 cells (Fig. 3B; Fig. S3B,D).

We also analysed possible alterations of chromosome conformation due to the CTCF site deletions with 3D-FISH using probes for *Shh*, ZRS and *Shh* brain enhancer 2 (SBE2), an enhancer that is located 460 kb upstream of the *Shh* coding sequence in the middle of the TAD (Jeong et al., 2006) (Figs 1A and 3D). Interprobe distances between all three probe pairs were significantly increased in CTCF deletion cells compared with wild-type ESCs (Fig. 3E; Table S2), consistent with the reduced interactions between *Shh* and

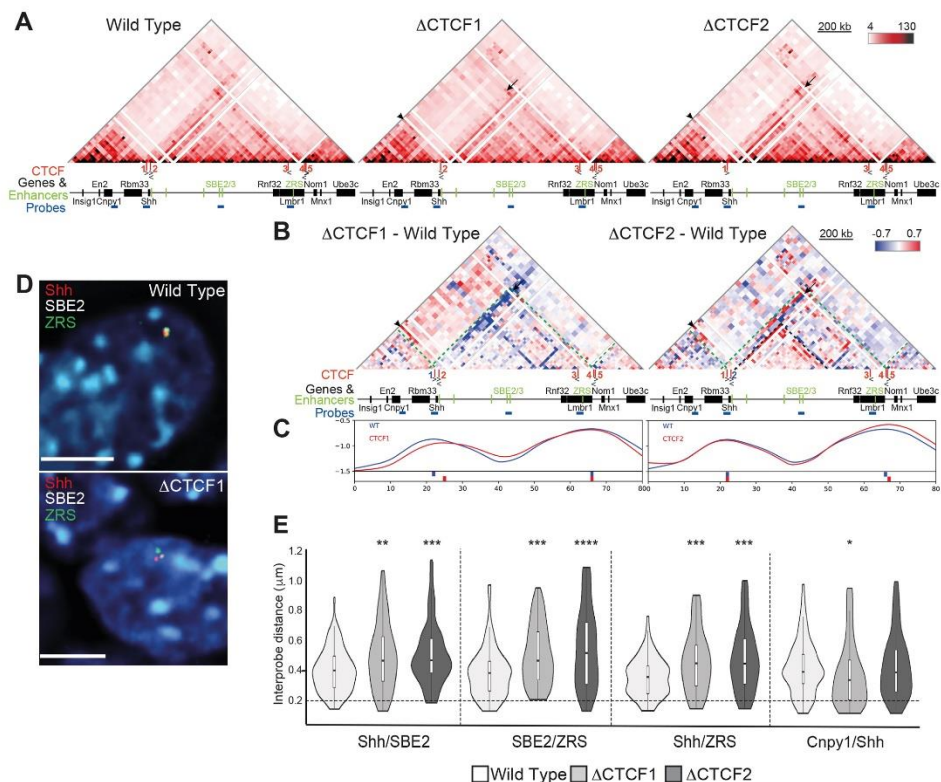


Fig. 3. 5C and 3D FISH identifies perturbations to chromatin conformation in Δ CTCF1 and Δ CTCF2 ESCs. (A,B) 5C heatmaps showing data from wild-type, Δ CTCF1 and Δ CTCF2 ESCs (A) and comparing Δ CTCF1 or Δ CTCF2 enrichment (red) with wild type (blue) (B). Green dashed lines indicate TAD boundaries in wild type, black arrows highlight loss of interactions in Δ CTCF1 or Δ CTCF2 cells and black arrowheads indicate enriched interactions in Δ CTCF1 or Δ CTCF2 cells. (C) Insulation score graphs that identify the location of TAD boundaries using raw summed 5C matrices. Peaks of insulation are shown below the plots, Δ CTCF1 or Δ CTCF2 (red) and wild type (blue). Data for biological replicates are in Fig. S3. (D) Images of representative nuclei from wild-type and Δ CTCF1 ESCs showing FISH signals for Shh/SBE2/ZRS probes. Scale bars: 5 μ m. (E) Violin plots showing the distribution of interprobe distances (μ m) between Shh/SBE2, SBE2/ZRS, Shh/ZRS and Cnpy1/Shh probes in wild-type, Δ CTCF1 and Δ CTCF2 ESCs. Horizontal dashed line shows the proportion of alleles that are colocalised (<200 nm). * <0.05 , ** <0.01 , *** <0.001 , **** <0.0001 (Mann–Whitney U-tests).

the rest of its TAD identified by 5C. Conversely, distances between *Shh* and *Cnpy1* (in the neighbouring *En2* TAD) were significantly decreased in Δ CTCF1 cells compared with wild type (Fig. 3E; Table S2), consistent with relocation of the TAD boundary.

Deleting either CTCF1 or CTCF2 disrupts *Shh*-ZRS spatial proximity in ESCs and, more generally, results in reduced 5C interactions between *Shh* and the rest of the regulatory TAD that may be due to the relocation of the TAD boundary (Δ CTCF1) or greater sub-division of the TAD (Δ CTCF2). The TAD boundary adjacent to *Shh* is sharply defined by CTCF1, whereas the boundary location of the neighbouring *En2* TAD cumulatively results from both CTCF1 and 2, possibly by blocking loop extrusion emanating from this TAD. However, neither of these deletions on their own is sufficient to cause merging of the two neighbouring TADs.

***Shh*-ZRS proximity is disrupted by the deletion of ZRS/*Lmbr1* CTCF sites**

Both CTCF1 and CTCF2 have enriched interactions with the CTCF site ~ 20 kb from ZRS in intron 5 of *Lmbr1* (CTCF3) (Fig. 1A). Therefore, we deleted both copies of CTCF3 (Δ CTCF3), described as i5 in Paliou et al. (2019).

Although whole TAD integrity was unaffected by Δ CTCF3 (Fig. 4A,C; Fig. S4A,C), intra-TAD reorganisation occurred in a manner similar to the loss of CTCF2, with enriched interactions within the sub-TADs (Fig. 4B; Fig. S4B,D). Loss of interactions between CTCF3 and *Shh*/CTCF2 in Δ CTCF3 cells appears to be somewhat compensated for by enriched contacts (arrowheads) between the ZRS locus and the *Shh* region of the TAD, particularly CTCF1 (Fig. 4B, inset heatmap adjacent to Δ CTCF3). Ectopic CTCF binding at ZRS has recently been identified following the

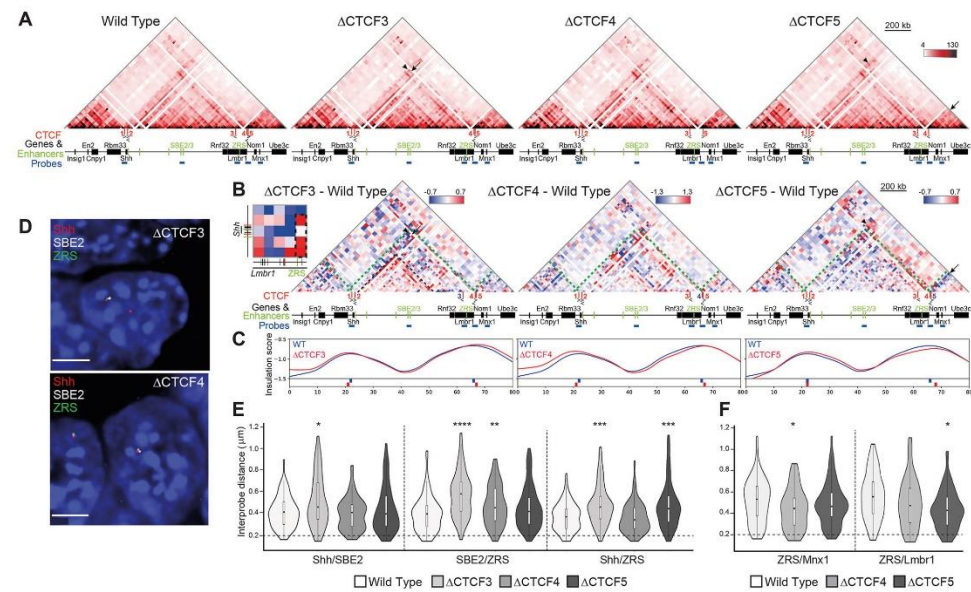


Fig. 4. 5C and 3D FISH identifies perturbations to local chromatin conformation in Δ CTCF3, Δ CTCF4 and Δ CTCF5 ESCs. (A,B) 5C heatmaps from wild-type, Δ CTCF3, Δ CTCF4 and Δ CTCF5 ESCs (A) and comparing Δ CTCF3, Δ CTCF4 or Δ CTCF5 enrichment (red) with wild type (blue) (B). Green dashed lines indicate the TAD boundaries in wild-type cells, black arrows highlight loss of interactions in Δ CTCF3 or Δ CTCF4 or Δ CTCF5 cells and black arrowheads indicate enriched interactions in Δ CTCF3 or Δ CTCF4 or Δ CTCF5 cells. Inset heatmap in B highlights enriched contacts between *Shh* and ZRS loci in Δ CTCF3; black dashed rectangle indicate interactions between ZRS bin and bins covering the *Shh* locus. (C) Insulation score graphs that identify the location of TAD boundaries using raw summed 5C matrices. Peaks of insulation are shown below the plots for Δ CTCF3, Δ CTCF4 or Δ CTCF5 (red) and wild type (blue). Data for biological replicates are in Fig. S4. (D) Images of representative nuclei from Δ CTCF3 and Δ CTCF4 ESCs showing FISH signals for *Shh*/SBE2/ZRS probes. Scale bars: 5 μ m. (E,F) Violin plots showing the distribution of interprobe distances (μ m) between *Shh*/SBE2, SBE2/ZRS and *Shh*/ZRS probes in wild-type, Δ CTCF3, Δ CTCF4 and Δ CTCF5 ESCs (E) and in ZRS/*Mnx1* and ZRS/*Lmbr1* probes in wild-type, Δ CTCF4 and Δ CTCF5 ESCs (F). Horizontal dashed line shows the proportion of alleles that are colocalised (<200 nm). * p <0.05, ** p <0.01, *** p <0.001, **** p <0.0001 (Mann–Whitney U -tests).

loss of neighbouring CTCF sites including CTCF3 (Paliou et al., 2019). Despite this compensation identified by 5C, FISH showed significantly increased interprobe distances between *Shh*, SBE2 and ZRS in Δ CTCF3 cells compared with wild type (Fig. 4D,E; Table S2). These data suggest that loss of any one of the three CTCF binding sites (1, 2 or 3) can disrupt the spatial proximity of *Shh*, SBE2 and ZRS (Figs 3E and 4E).

Finally, we generated ESC lines with deletions of CTCF binding sites at the *Lmbr1* promoter (Δ CTCF4) and 5' *Lmbr1* (Δ CTCF5), both of which are located at the boundary between the *Shh* TAD and the adjacent TAD containing *Mnx1* (Fig. 1A). The CTCF motif within CTCF5 is oriented towards the *Mnx1*-containing TAD and deletion of CTCF5 caused a loss of interactions between this boundary region and the *Mnx1* TAD (Fig. 4A,B; Fig. S4B,D) with the TAD boundary shifted toward *Nom1* (Fig. 4C). 5C reveals increased interactions in the ZRS-*Lmbr1* region in Δ CTCF5 (Fig. 4B) and FISH also shows increased spatial proximity between ZRS and *Lmbr1* (Fig. 4F; Table S2).

FISH revealed significantly increased interprobe distances between *Shh* and ZRS in Δ CTCF5 cells and between SBE2 and ZRS in Δ CTCF4 cells, which is not detected by 5C (Fig. 4E). There are also decreased distances seen between ZRS and *Mnx1* in the

adjacent TAD in the absence of *Shh*-CTCF4, also not apparent in the 5C data (Fig. 4F).

We conclude that deletion of CTCF binding sites at either of the *Shh* TAD and sub-TAD boundaries, especially CTCF1, 2 and 3, affects local chromatin organisation in ESCs and disrupts *Shh*/ZRS spatial proximity.

Reduced *Shh*-ZRS colocalisation in the limb upon the loss of CTCF1, 2 and 3

To test how disrupted TAD organisation impacts on chromosome conformation and *Shh* gene expression during embryonic development, we generated mouse lines carrying each of the homozygous CTCF deletions. We have previously reported enhanced *Shh*-ZRS colocalisation in the limb bud at the time and place of *Shh* expression that depends on a fully functional ZRS (Lettice et al., 2014; Williamson et al., 2016). Therefore, we assayed the spatial proximity of *Shh*, SBE2 and ZRS by FISH in E11.5 embryo sections that include posterior (ZPA) and anterior distal limb tissue from wild-type and homozygous Δ CTCF mutant embryos (Fig. 5A).

In both regions of the wild-type limb bud analysed (ZPA and anterior), *Shh*-ZRS distances were shorter than between either *Shh*-SBE2 or SBE2-ZRS, consistent with *Shh* and ZRS being

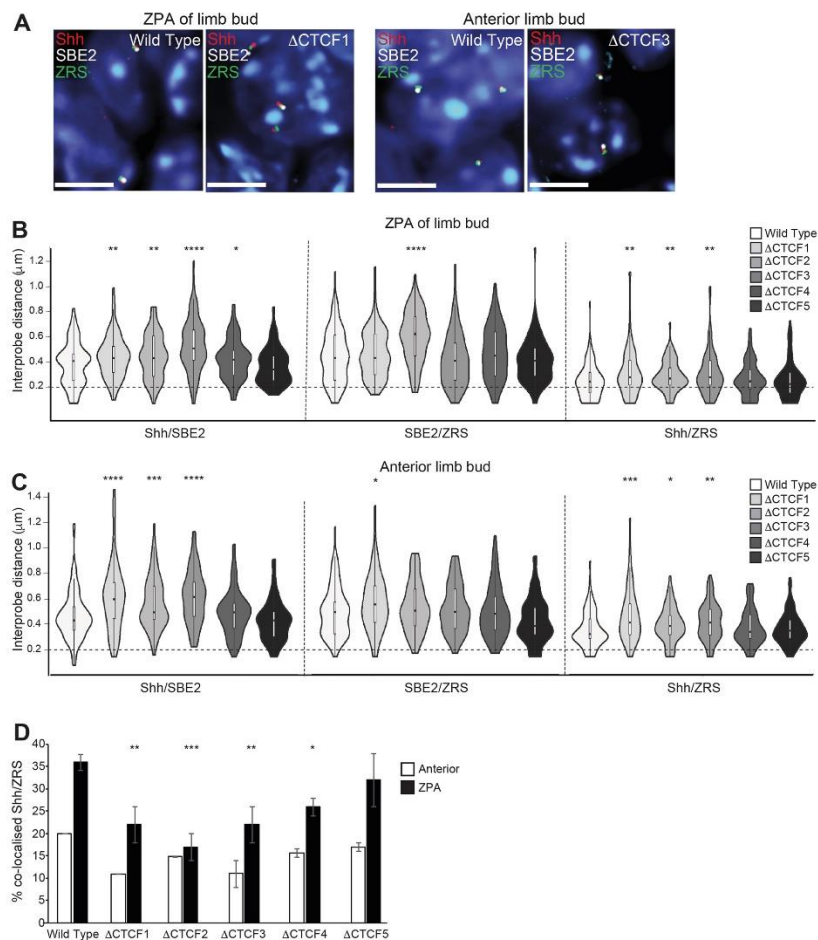


Fig. 5. Perturbation of chromatin conformation within the *Shh* TAD in the distal limb bud of Δ CTCF mutant embryos. (A) Images of representative nuclei from E11.5 ZPA and distal anterior limb bud in wild-type, Δ CTCF1 and Δ CTCF3 embryos showing FISH signals for *Shh*/*SBE2*/*ZRS* probes. Scale bars: 5 μ m. (B,C) Violin plots showing the distribution of interprobe distances (μ m) between *Shh*/*SBE2*, *SBE2*/*ZRS* and *Shh*/*ZRS* probes in E11.5 wild type and Δ CTCF mutant embryos in ZPA/distal posterior (B) and distal anterior limb bud (C). Horizontal dashed lines show the proportion of alleles that are colocalised (<200 nm). (D) Histograms showing the percentage of colocalised *Shh*/*ZRS* probe pairs (<200 nm) in wild-type and each of the Δ CTCF E11.5 embryos for distal anterior and ZPA limb bud tissue. Error bars represent s.e.m. obtained from two or three different tissue sections. * <0.05 , ** <0.01 , *** <0.001 , **** <0.0001 (Mann-Whitney *U*-tests in B,C; Fisher's Exact Tests in D).

maintained in spatial proximity across the limb bud (Fig. 5B,C). Similar to our observation in ESCs (Figs 3 and 4), distances between *Shh* and both *SBE2* and *ZRS* were significantly increased in Δ CTCF1, 2 and 3, but not in Δ CTCF4 and 5 embryos (Fig. 5B,C; Tables S3 and S4). The frequency of *Shh*-*ZRS* colocalisation (<200 nm) in the ZPA of Δ CTCF1, 2, and 3 mutant embryos was reduced to levels seen in non-expressing parts of the wild-type limb bud (Fig. 5D; Table S5).

***Shh* expression patterns and development are unaffected in CTCF site-deletion mice**

Our data indicate that deletion of individual CTCF sites can affect TAD boundaries, intra- and inter-TAD interactions and enhancer-promoter colocalisation frequencies. These alterations in 3D chromosome conformation might be predicted to affect gene expression. However, we found that mice homozygous for any of the Δ CTCF deletions are viable, fertile and have no overt deleterious

phenotype. *In situ* hybridisation in homozygous mutant embryos showed a normal pattern of *Shh* expression in the brain (Fig. 6A), and body (Fig. 6B) at similar levels to wild type. At E11.5 expression is detected only within the developing midline of the brain, the zona limitans intrathalamica and the medial ganglionic eminence in the head, and staining is visible in the floor plate and notochord, the zone of polarising activity (ZPA) of the limb buds and umbilicus in the body. No ectopic expression is detected at the midbrain/hindbrain junction driven by neighbouring *En2* or *Cnpy1* enhancers (Fig. 6B). Conversely, in embryos homozygous for either Δ CTCF1 or Δ CTCF2 there is no evidence for ectopic *En2* and *Cnpy1* expression in any of the normal sites of *Shh* expression in the brain (Fig. 6C,D).

Similarly, no ectopic *Shh* expression is detected in motor neurons driven by *Mnx1* enhancers in the TAD beyond ZRS/*Lmbr1* (Lee et al., 2004) (Fig. 6A), and *Mnx1* was not expressed ectopically in any of the normal sites of *Shh* expression in embryos carrying homozygous deletions of CTCF3, 4 or 5 (Fig. 6E). These findings indicate that despite the alterations to *Shh* TAD architecture and chromosome conformation, enhancer/promoter specificity is maintained in the Δ CTCF embryos and that, in the absence of these CTCF sites, there is no crosstalk across TAD boundaries to result in ectopic expression driven by *Shh* enhancers.

In situ hybridisation is good for determining spatial expression patterns but is at best a semi-quantitative technique. Therefore, we used RNA FISH to detect nascent *Shh* transcripts in regions of the developing brain in order to quantify the number of expressing alleles in individual cells (representative images are shown in Fig. S5A). In a region where *Shh* expression is driven by SBE2 (Z. Crane-Smith, personal communication) we detected a small but significant reduction in the percentage of *Shh* expressing alleles in embryos carrying homozygous deletions of CTCF2, 3 and 5 (Fig. 6F). However, when we used qRT-PCR to examine RNA from entire heads, in which expression is controlled by multiple enhancers, no significant changes in *Shh* mRNA levels were observed (Fig. 6G).

Shh expression in the limb bud driven by ZRS lasts only 48 h from initiation to downregulation. qRT-PCR performed from limb buds at E10.5 can detect changes in expression in response to deletions within the *Shh* TAD (Paliou et al., 2019). However, we detected no differences in the percentage of expressing *Shh* alleles on E11.5 developing limb buds using RNA FISH (Fig. S5B), and qRT-PCR shows no significant differences in *Shh* mRNA expression levels, with the exception of Δ CTCF5 mutants (Fig. S5C).

Mice heterozygous for a *Shh* null allele express only 50–60% wild-type levels of *Shh* in the limb bud but develop normally; in fact, in the limb *Shh* levels must fall to ~20% of wild type before development is perturbed and digits are lost (Lettice et al., 2017). As the TAD boundary moves beyond the 5' end of *Shh* in Δ CTCF1 cells, arguably separating the coding region from its enhancers, we also made compound heterozygotes carrying both the Δ CTCF1 and *Shh* null alleles to uncover subtle effects on *Shh* expression. These *Shh* ^{Δ CTCF1/-} mice develop normally and are viable and fertile, further suggesting that deletion of CTCF1 results in no deleterious changes in *Shh* expression.

A 35 kb deletion that removes the *Lmbr1* promoter and TAD boundary disrupts chromatin conformation with no deleterious phenotype

Deletion of CTCF1 3' of *Shh* showed that this position was important for the TAD boundary location and for *Shh* physical proximity with its regulatory domain (Fig. 3; Fig. S3), but the loss of

this site had no apparent phenotypic consequence (Fig. 6). In a similar manner, deleting CTCF5 at the *Lmbr1* TAD boundary affected the *Shh* TAD boundary location but resulted in a minimal loss of proximity between ZRS and *Shh*, at least in E11.5 limb tissue (Fig. 4; Figs S4 and S5). Loss of TAD boundary regions can result in the merging of adjacent TADs and the ectopic activation of genes in one TAD by enhancers in the other merged TAD, with phenotypic consequences (Fabre et al., 2017; Lupiáñez et al., 2015). However, this involved the deletion of sizeable stretches of DNA across the boundaries in question, tens of kilobases rather than individual CTCF sites. In addition to CTCF binding sites, a number of features are found enriched at TAD boundaries, including those associated with active promoters (Dixon et al., 2012). To determine whether a more extensive deletion across the *Lmbr1* boundary results in the merging of adjacent TADs, a homozygous 35 kb deletion (Δ 35) was generated in mice which removed CTCF4 and CTCF5 and covering a region containing the first two exons of *Lmbr1* and 13 kb upstream (Fig. 1A). RT-PCR in mouse embryos confirmed that this deletion eliminates transcription throughout the 5' end of *Lmbr1* in both isolated limb buds and the rest of the body (Fig. S6A).

5C from homozygous Δ 35 ES cells showed that this deletion caused relocation of the TAD boundary a further ~60 kb 5' of the *Lmbr1* promoter towards the promoter of *Nom1* rather than a merging of the adjacent TADs (Fig. 7A; Fig. S6B,C). Interactions within the region extending from CTCF3 to *Nom1* are enriched in Δ 35 cells compared with wild type, and the CTCF3/ZRS genomic region gains interactions into the adjacent TAD up to *Mnx1* (Fig. 7A, arrowheads). *Nom1* loses interactions within its own TAD (Fig. 7A, arrows; Fig. S6B,C). Consistent with this, 3D-FISH (Fig. 7B) showed that distances between ZRS and *Shh*, SBE2 and *Mnx1* were all significantly decreased in Δ 35 (Fig. 7C; Table S6). The reduced spatial distance between ZRS and *Mnx1* was not due to reduction of the linear genomic distance caused by the 35 kb deletion, as similar effects were seen in cells carrying an inversion of this 35 kb of DNA (Fig. 7C).

Deletions of the *Shh* TAD boundary at the *Lmbr1* promoter relocates the boundary to the promoter of *Nom1*, and the ZRS has enhanced ability to contact sequences both within its own TAD and the *Mnx1* TAD. Despite these differences, Δ 35 homozygous mice were viable, fertile and had no apparent phenotype. The *Shh* expression pattern is also indistinguishable from wild type (Fig. 7D, E) – in particular midline expression is detected in the floor plate and notochord as one stripe down the body (Fig. 7E, arrowhead), with no evidence for expression as two more lateral stripes driven by *Mnx1* motor neuron enhancers (Lee et al., 2004; Fig. 6E). Because of the changes in interactions across the TAD boundary observed in Δ 35, we made compound heterozygotes with the *Shh* null chromosome to highlight subtle changes in expression. Even in this sensitised *Shh* background, compound *Shh* ^{Δ 35/-} mice are phenotypically normal.

Interestingly, given the decreased distances measured by FISH between ZRS and *Mnx1*, *in situ* hybridisations indicate that the limb expression of *Mnx1* is increased in *Shh* ^{Δ 35/ Δ 35} embryos in comparison with wild-type embryos (Fig. 7F–H). This is supported by qRT-PCR which detects a modest, but not significant, upregulation of *Mnx1* expression in limb buds, whereas *Shh* expression is unchanged (Fig. 7I). These data suggest that the deletion of 35 kb encompassing the TAD boundary enhances the ability of the *Mnx1* promoter to respond to the ZRS. However, no upregulation of *Mnx1* expression is seen in the pharyngeal endoderm and developing lungs, which would be driven by the enhancers neighbouring ZRS, MACS1 and MFCS4 (Fig. 1A) (Sagai et al., 2009).

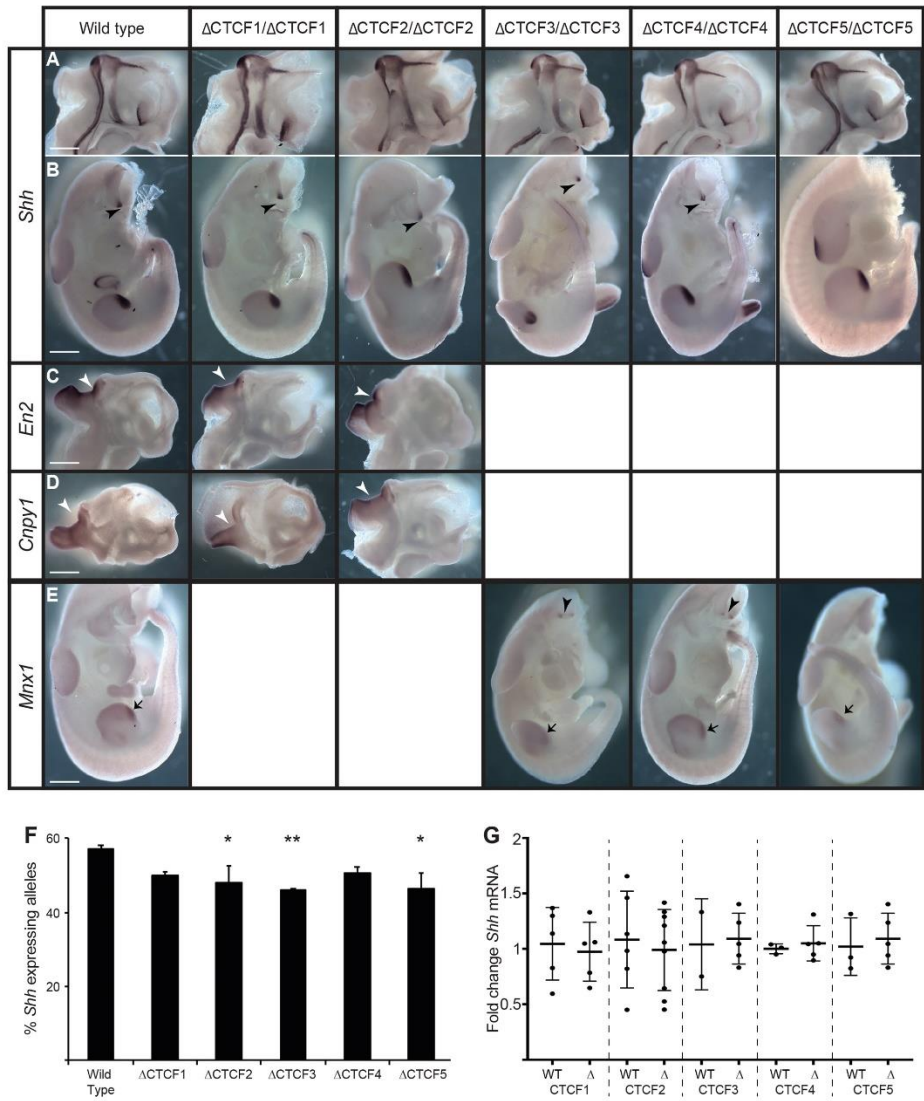


Fig. 6. Expression patterns of *Shh* and genes in neighbouring TADs are unaffected by CTCF site deletions. (A–E) *In situ* analysis of gene expression at E11.5 in wild type and embryos homozygous for each of the Δ CTCF lines. (A,B) Normal expression of *Shh* in the midline of bisected heads (A) and in the bodies (B) where expression is detected in the ZPA of the limb bud and in the floorplate and notochord (arrowheads). (C,D) Expression of *En2* (C) or *Cnpy1* (D) in bisected heads. Expression is detected only at the mid brain hindbrain junction (arrowheads). (E) Expression of *Mnx1* in the ZPAs of the limb buds (arrows) and in the motor neurons (arrowheads). Scale bars: 1 mm. (F) Percentage of expressing *Shh* alleles detected by RNA FISH within the region of the brain in which expression is controlled by SBE2. * <0.05 , ** <0.01 (Fisher's Exact Tests). Error bars represent s.e.m. (G) Fold change of *Shh* expression determined by qRT-PCR of cDNA made from E11.5 heads. Mutants from each of Δ CTCF lines are compared with wild-type litter mates. Each dot represents a single embryo and data are means $\pm$ s.d. (unpaired Student's *t*-tests).

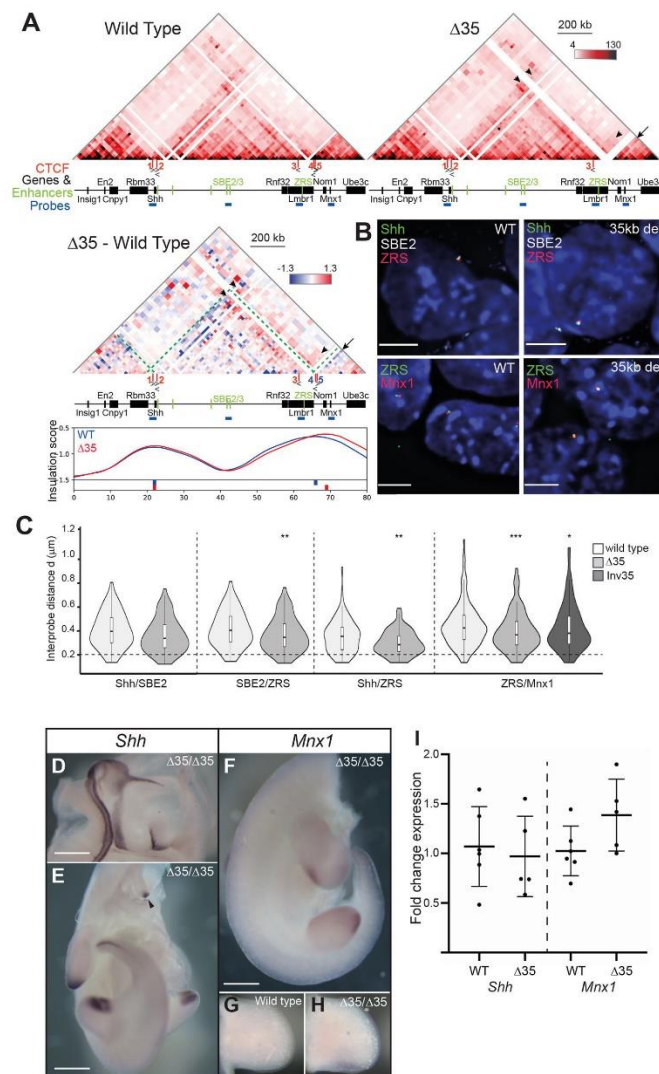


Fig. 7. Chromosome conformation and gene expression as a consequence of a 35-kb deletion ($\Delta 35$) at the *Lmbr1* TAD boundary. (A) 5C heatmaps from wild-type and $\Delta 35$ ESCs. Below shows a heatmap comparing $\Delta 35$ 5C enrichment (red) with wild type (blue). Green dashed lines indicate the *Shh* TAD boundaries, black arrows highlight loss of interactions in $\Delta 35$ cells and black arrowheads indicate enriched interactions in $\Delta 35$ cells. At the bottom are insulation score graphs that identify the location of TAD boundaries using raw summed 5C matrices. Peaks of insulation are shown below the plots, $\Delta 35$ (red) and wild type (blue). Data for biological replicates are in Fig. S6. (B) Images of representative nuclei from wild-type and $\Delta 35$ ESCs showing FISH signals for *Shh*/SBE2/ZRS and ZRS/*Mnx1* probes. Scale bars: 5 μ m. (C) Violin plots show the distribution of interprobe distances (μ m) between *Shh*/SBE2, SBE2/ZRS, *Shh*/ZRS and ZRS/*Mnx1* probes in wild-type and $\Delta 35$ ESCs, and ZRS/*Mnx1* distances in 35 kb inversion ESCs. Horizontal dashed line shows the proportion of alleles that are colocalised (<200 nm). * p <0.05, ** p <0.01, *** p <0.001 (Mann–Whitney *U*-Tests). (D,E) *In situ* hybridisations showing normal *Shh* expression in a bisected head and body, respectively, of a $\Delta 35/\Delta 35$ E11.5 embryo. (F–H) *In situ* hybridisations for *Mnx1* in a $\Delta 35/\Delta 35$ homozygote (F) and limb bud (H) and, for comparison, lower levels of staining in a wild-type limb bud is shown in (G). The staining in G and H was stopped before wild-type signal was apparent to highlight *Mnx1* upregulation. Scale bars: 1 mm. (I) Fold change of *Shh* and *Mnx1* expression determined by qRT-PCR of cDNA made from E11.5 limb buds. Mutant $\Delta 35/\Delta 35$ embryos are compared with wild-type litter mates. Each dot represents a single embryo and data are means $\pm$ s.d. (unpaired Student's *t*-tests).

DISCUSSION

A systematic genetic approach to delete individual CTCF sites, and to delete or invert large regions, including those encompassing a TAD boundary, has enabled us to use chromosome conformation capture and imaging to assay the resulting perturbations to chromosome organisation within the *Shh* regulatory TAD, and between this and neighbouring TADs. Analysing CTCF deletions, we detected little or no disruption to gene regulation during

embryonic development, and no detectable phenotype in animals that can be attributed to these altered chromosome conformations.

ZRS activity is not distance dependent

5C analysis confirmed that TAD boundaries were unaffected by removal of most of the internal region of the *Shh* TAD ($\Delta 700$) (Fig. 1; Fig. S1), with *Shh* and its remaining enhancers still located

within the same, but smaller, TAD. This large deletion did cause extensive disruption to the developing embryo, mainly, it can be assumed, due to the loss of several known forebrain and epithelial enhancers within the deleted region. However, even in embryos homozygous for the 700 kb deletion, which relocates ZRS to less than 100 kb distant from *Shh*, ZRS function is maintained, with no detrimental effects on limb bud-specific *Shh* activation, and normal development of the limbs occurs. Therefore, the large genomic distance from *Shh* is not intrinsic to the function of the ZRS. This is in contrast to the loss of interactions following similar perturbations between a limb-specific enhancer and *Hoxd13* that resulted in loss of *Hoxd13* activity (Fabre et al., 2017).

Loss of CTCF sites at the *Shh* TAD boundaries disrupts chromatin architecture and impacts *Shh*/ZRS spatial proximity

We have previously shown that *Shh* and ZRS are in spatial proximity (~300 nm) in the early embryo in both expressing limb tissue and the non-expressing adjacent flank (Williamson et al., 2016). Here, using 5C on cells dissected from E11.5 anterior and posterior limb buds, we show that this is driven by an interaction between the sites 3' and 5' of *Shh* (containing CTCF1 and CTCF2 sites) and a region within intron 5 of *Lmbr1* ~20 kb from ZRS (CTCF3) (Fig. 2; Fig. S2). This loop is also present in ESCs, and spatial proximity of *Shh* and ZRS is lost upon the deletion of any one of the three CTCF sites in both ESCs and E11.5 limb bud tissue (Figs 3-5; Figs S3 and S4). Deleting CTCF sites at the *Lmbr1* promoter TAD boundary (Δ CTCF4 and Δ CTCF5) had a lesser effect on *Shh*/ZRS spatial proximity. Increased interprobe distances between either *Shh* or ZRS and the forebrain enhancer SBE2 located at the centre of the TAD suggest that the loss of spatial proximity may be due to a general decompaction throughout the TAD, rather than a loss of interactions which could be detected by 5C.

Shh responds to its developmental enhancers regardless of TAD disruption

5C analysis in ESCs suggests that the disruption caused by CTCF site deletions can remove *Shh* from its regulatory TAD (Δ CTCF1) or re-enforce contacts within sub-TAD domains such that the *Shh* forebrain enhancers are sequestered in one, and ZRS and the long-range epithelial enhancers are sequestered in the other, with a loss of interactions between both sub-TADs and either sub-TAD with *Shh* (Δ CTCF2 and Δ CTCF3). Nevertheless, in all of these configurations, the expression pattern of *Shh* during embryonic development appears to be normal and mRNA levels are largely unchanged, with the resulting mice having no detectable phenotype. This indicates that communication between *Shh* and its extensive set of developmental enhancers is remarkably robust to TAD perturbation.

Ectopic expression across disrupted TAD boundaries is not common

Loss of CTCF1 not only moves the TAD boundary ~60 kb to beyond the 5' end of *Shh* but also enables greater interactions between *Shh* and the adjacent TAD, which contains other genes and their enhancers that are active during brain development, but in a pattern distinct from *Shh*. *En2* is expressed at the mid-hindbrain boundary, a pattern at least partly dependent on an enhancer binding Pax2/5/8 (Li Song and Joyner, 2000). Similarly, *Cnpy1* expression at the mid-hindbrain boundary is thought to be important for FGF signalling (Hirate and Okamoto, 2006). Despite increased chromatin interactions over the *Shh* TAD boundary in Δ CTCF1,

there is no ectopic expression of *Shh* in the mid-hindbrain driven by the *En2/Cnpy1* enhancers and, vice versa, there is no ectopic expression of *En2/Cnpy1* at sites driven by *Shh* enhancers (Fig. 6).

The *Lmbr1* TAD boundary has been suggested to be less precise than that at the *Shh* end of the TAD from both a structural and regulatory point of view (Anderson et al., 2014; Symmons et al., 2016). Deletion of CTCF5 weakened the boundary of the neighbouring *Mnx1* TAD, and increased proximity between ZRS and *Mnx1* was detected in Δ CTCF4. However, in neither case was there evidence for enhanced expression of *Mnx1* – e.g. in limb buds driven by ZRS – beyond that detected in wild-type embryos. Interestingly, even in wild-type situations, *Mnx1* has a weak expression domain concomitant with the limb bud ZPA, suggesting that this gene may be influenced by ZRS activity emanating from the adjacent TAD. Nor was there evidence of the *Mnx1* motor-neuron enhancer (Zelenchuk and Bruses, 2011) driving expression of *Shh* in motor neurons of the developing neural tube in any of the mutant embryos.

A larger (35 kb) deletion of this boundary, removing CTCF4, CTCF5 and the promoter/first two exons of *Lmbr1*, enhanced ZRS contacts across both the *Shh* TAD and into the neighbouring *Mnx1* TAD (Fig. 7; Fig. S6). Increased *Mnx1* expression in the ZPA of embryos homozygous for the 35 kb deletion suggests that the potentially increased contacts between *Mnx1* and ZRS identified in ESCs could be enabling greater activation of this gene by the *Shh* limb enhancer.

Perturbations of the *Shh* TAD boundaries can negatively impact on gene-enhancer colocalisation but are insufficient to cause a deleterious phenotype

It is commonly held that enhancer-driven gene activation requires 'contact' or very close apposition of the enhancer and promoter. Inversions encompassing the *Shh* TAD boundaries that disrupted TAD integrity and significantly increased the genomic distance between *Shh* and ZRS result in severe limb malformations, suggesting that these rearrangements prevent ZRS from contacting/regulating the *Shh* promoter (Symmons et al., 2016). These data and our 5C and FISH analyses, which show that the *Shh* TAD forms a compact discrete regulatory hub (Williamson et al., 2016), suggest that 3D organisation of the *Shh* TAD could allow distal enhancers to come into close proximity to selectively regulate *Shh* expression. However, in the functionally relevant cells of the limb bud ZPA, ZRS colocalisation (<200 nm) with *Shh* was reduced to levels of the non-expressing distal anterior levels in Δ CTCF1, Δ CTCF2 and Δ CTCF3 homozygous embryos without adversely affecting *Shh* expression (Fig. 6) and with no subsequent phenotypic effects. This is consistent with evidence showing reduced *Shh* neural enhancer-promoter colocalisation in expressing cells and tissues (Benabdallah et al., 2019).

All embryos homozygous for one of the five CTCF binding domain deletions or the 35 kb deletion of the *Lmbr1* boundary developed normally and were able to reproduce. Moreover, sufficient *Shh* expression was maintained for compound heterozygote embryos carrying either Δ CTCF1 or the 35 kb deletion opposite a *Shh* null allele to have no abnormal phenotype. A contemporaneous study on the same genomic territory has largely recapitulated these results – deletions of *Lmbr1* CTCF sites and the gene promoter caused perturbations to local chromatin conformation but *Shh* expression, although reported to be reduced, was enough to drive normal limb development (Paliou et al., 2019). The *Shh* regulatory landscape is set up to ensure optimal activation of the gene and here, we have shown that

this is robust to perturbations of TAD integrity and structure. Similarly, recent work on the *Sox9-Kcnj2* locus suggests that even manipulations that result in the fusion of neighbouring TADs have no major effects on gene expression (Despang et al., 2019). However, large-scale disruptions incorporating boundaries which cause TADs to merge do result in developmental defects (Lupiáñez et al., 2015).

Our data suggest that CTCF binding has a role in TAD structure and loss of sites perturbs internal interactions and the position of boundaries. However, at the *Shh* locus these major disruptions have no effect on gene expression patterns and little effect on expression levels. We speculate that the largely unvarying organisation of TADs could have provided the necessary stable genomic environment for the accumulation of regulatory elements over evolutionary time rather than being essential for target gene activation.

MATERIALS AND METHODS

Cell culture and CRISPR-Cas9 mediated deletions

E14TG2A mouse ESCs (a kind gift from Austin Smith, Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, UK) were cultured under standard conditions (Anderson et al., 2014). CRISPR guides were made by cloning annealed oligos (Table S1) into pSpCas9(BB)-2A-GFP (PX458, Addgene plasmid #48138), a gift from Feng Zhang (Ran et al., 2013). We transfected 2 µg of vector DNA into 8×10^5 ESCs using Lipofectamine 2000 (Thermo Fisher Scientific) following the manufacturer's instructions. After 48 h, GFP-positive cells were sorted using fluorescence-activated cell sorting and plated at low density. Ten days later, individual clones were picked and screened for correct deletion using PCR and Sanger sequencing (primers are listed in Table S1).

Mouse lines and embryo analysis

The *Shh*^{Δ700} deletion was created by crossing the line SBLac96 (Anderson et al., 2014) to a line carrying a pCAGGS-Cre recombinase gene (Araki et al., 2006). With the exception of the Δ35 kb and Inv35 kb mouse lines, which were made by injection of the ESCs into blastocysts, all of the other mouse lines were created as in Lettice et al. (2017) by direct microinjection into C57Bl6/CBA F2 zygotes of the same guides as were used in ESCs. Resultant G0 mice were screened by PCR using flanking primers (Table S1) and the deletions confirmed using Sanger sequencing. Lines were then established by crossing founder mice to C57Bl6 wild types. *lacZ* expression analysis, *in situ* hybridisations and RT-PCR reactions were conducted as in Anderson et al. (2014).

All mouse work has been approved by the University of Edinburgh Animal Welfare and Ethics Review board and is conducted under the authority of Home Office Licences.

qRT-PCR

Heads and limb buds were dissected from individual E11.5 embryos, snap frozen in separate tubes and stored at -80°C . The rest of the embryo was used to make DNA, and the genotype established by PCR using the primers listed in Table S1. RNA was extracted from the tissues of mutant and wild-type embryos using TRIzol (Invitrogen) and first strand cDNA synthesised with a Transcriptor First strand cDNA synthesis kit (Roche) following the manufacturers' instructions. qRT-PCR was run on an LC480 lightcycler (Roche) and made use of the Universal ProbeLibrary (probe 32 for *Shh*, probe 60 for *Mnx1* and the Universal ProbeLibrary Mouse GAPD Gene Assay). PCR primers are listed in Table S7. Each gene was assayed in separate wells and each sample run in triplicate. Gene expression data was analysed by the $\Delta\Delta\text{Ct}$ method, with mutants from each line compared with their wild-type littermates. ($n=3-9$ embryos). Student's unpaired *t*-test was used for statistical validations.

DNA FISH

E11.5 embryos were collected, fixed, embedded, sectioned, antibody stained for SHH expression and processed for FISH as previously described

(Morey et al., 2007; Lettice et al., 2014), except that sections were cut at 8 µm. Regions expressing Shh were identified by antibody staining with an anti-Shh antibody (Ab86462, Batch GR182460-5, Abcam; 1:150), which has been shown to give the correct expression pattern in immunostaining (<https://www.abcam.com/sonic-hedgehog-antibody-m0128-4a37-ab86462.html>). Fosmid clones (Fig. 1A; Table S8) were prepared and labelled as previously described (Morey et al., 2007). Between 160-240 ng of biotin- and digoxigenin-labelled fosmid probes were used per slide, with 16-24 µg of mouse Cot1 DNA (Invitrogen) and 10 µg salmon sperm DNA. For four-colour FISH, a similar quantity of the additional fosmid was labelled with either Green496-dUTP (Enzo Life Sciences) or red-dUTP (Alexa Fluor 594-5-dUTP, Invitrogen).

For 3D FISH on ESCs, 1×10^6 cells were seeded on slides overnight. Cells were fixed in 4% paraformaldehyde for 10 min at room temperature and then permeabilised using 0.5% Triton X for 10 min (Eskeland et al., 2010).

RNA FISH

Custom Stellaris® RNA FISH Probes were designed against *Shh* nascent mRNA (pool of 48 unique 22-mer probes) using the Stellaris® RNA FISH Probe Designer (Biosearch Technologies; www.biosearchtech.com/stellarisdesigner, version 4.2). The slides were hybridised with the *Shh* Stellaris FISH Probe set labelled with Quasar 570 (Biosearch Technologies), following the manufacturer's instructions (www.biosearchtech.com/stellarisprotocols). Briefly, FFPE tissue sections from E11.5 embryos were deparaffinised in xylene, hydrated in ethanol and permeabilised in 70% ethanol overnight at 4°C . Slides were incubated in 10 µg/ml proteinase K in 1× PBS for 20 min at 37°C followed by washes in 1× PBS and wash buffer (2× SSC, 10% deionised formamide). *Shh* RNA FISH probes were diluted in Stellaris RNA FISH hybridisation buffer (#SMF-HB1-10) to 125 nM and hybridised to slides overnight in a humidified chamber at 37°C . Slides were washed twice for 30 min in wash buffer (2× SSC, 10% deionised formamide) at 37°C , counterstained with 5 ng/ml DAPI, washed in 1× PBS and mounted in Vectashield (Vector Laboratories).

Image analysis

Slides were imaged using a Photometrics Coolsnap HQ2 CCD camera and a Zeiss AxioImager A1 fluorescence microscope with a Plan Apochromat 100×1.4NA objective, a Nikon Intensilight Mercury based light source and either Chroma #89014ET (three-colour) or #89000ET (four-colour) single excitation and emission filters (Chroma Technology Corp.) with the excitation and emission filters installed in Prior motorised filter wheels. A piezoelectrically driven objective mount (PIFOC model P-721, Physik Instrumente) was used to control movement in the *z* dimension. Step size for *z* stacks was set at 0.2 µm. Hardware control, image capture and analysis were performed using Nikon Nis-Elements software. Images were deconvolved using a calculated point spread function with the constrained iterative algorithm of Volocity (PerkinElmer). Image analysis was carried out using the Quantitation module of Volocity. For DNA FISH, only alleles with single probe signals were analysed, to eliminate the possibility of measuring sister chromatids.

3C library preparation

Limbs buds and bodies (with the limbs and heads removed) from wild-type embryos, and entire *Shh*^{Δ700/Δ700} embryos were dissected at E11.5 and the tissue dissociated by pipetting in just enough PBS to cover them. The cells were fixed with 1% formaldehyde for 10 min at room temperature. For ESCs, 5×10^6 – 1×10^7 cells were fixed. Crosslinking was stopped with 125 mM glycine for 5 min at room temperature followed by 15 min on ice. Cells were centrifuged at 400 *g* for 10 min at 4°C , supernatants removed, and cell pellets flash frozen on dry ice before storage at -80°C .

Cell pellets were treated as previously described (Dostie and Dekker, 2007; Ferraiuolo et al., 2010; Williamson et al., 2014). HindIII-HF (New England Biolabs) was the restriction enzyme used to digest the crosslinked DNA.

5C primer and library design

5C primers covering the *Usp22* (mm9, chr11: 60,917,307-61,003,268) and *Shh* regions (mm9, chr5: 28,317,087-30,005,000) were designed using

'my5C.primer' (Lajoie et al., 2009) with the following parameters: optimal primer length of 30 nt, optimal primer melting temperature (TM) of 65°C, default primer quality parameters (mer:800, U-blast:3, S-blast:50). Primers were not designed for large (>20 kb) and small (<100 bp) restriction fragments, for low complexity and repetitive sequences, or where there were sequence matches to >1 genomic target. The *Usp22* region was used to assess the success of each 5C experiment but was not used for further data normalisation or quantification.

The universal A-key [CCATCTCATCCCTGCGTGTCTCCGACTCAG- (5C-specific)] and the P1-key tails [(5C-specific)-ATCACCGACTGCC-ATAGAGAGG] were added to the forward and reverse 5C primers, respectively. Reverse 5C primers were phosphorylated at their 5' ends. An alternating design consisting of 365 primers in the *Shh* region (182 forward and 183 reverse primers) was used. Primer sequences are listed in Table S9.

5C library preparation

5C libraries were prepared and amplified with the A-key and P1-key primers as described in Fraser et al. (2012). Briefly, 3C libraries were first titrated by PCR for quality control (single band, absence of primer dimers, etc.), and to verify that contacts were amplified at frequencies similar to those usually obtained from comparable libraries (same DNA amount from the same species and karyotype) (Dostie and Dekker, 2007; Dostie et al., 2007; Fraser et al., 2010). We used 1–10 µg of 3C library per 5C ligation reaction.

5C primer stocks (20 µM) were diluted individually in water on ice and mixed to a final concentration of 2 nM. Mixed diluted primers (1.7 µl) were combined with 1 µl of annealing buffer (10× NEBuffer 4, New England Biolabs) on ice in reaction tubes. Then 1.5 µg salmon testis DNA was added to each tube, followed by the 3C libraries and water to a final volume of 10 µl. Samples were denatured at 95°C for 5 min and annealed at 55°C (48°C for ESCs) for 16 h. Ligation with Taq DNA ligase (10 U) was performed at 55°C (48°C for ESCs) for one hour. One tenth (3 µl) of each ligation was then PCR-amplified individually with primers against the A-key and P1-key primer tails. We used 26 cycles based on dilution series showing linear PCR amplification within that cycle range. The products from three to five PCR reactions were pooled before purifying the DNA on MinElute columns (Qiagen).

5C libraries were quantified by bioanalyser (Agilent) and diluted to 26 pmol (for Ion PGM Sequencing 200 Kit v2.0). We then used 1 µl of diluted 5C library for sequencing with an Ion PGM Sequencer. Samples were sequenced onto Ion 316 Chips following the Ion PGM Sequencing 200 Kit v2.0 protocols as recommended by the manufacturer (Life Technologies).

5C data analysis

Analysis of the 5C sequencing data was performed as described in Berlivet et al. (2013). The sequencing data were processed through a Torrent 5C data transformation pipeline on Galaxy (<https://main.g2.bx.psu.edu/>). Before normalising, interactions between adjacent fragments were removed owing to the high noise:signal ratio likely to occur here. Average read count values over 21 kb bins were calculated from the raw sequencing data and 5C data were further processed for visualisation. First, the matrices were normalised to sum up to 50,000 reads (excluding the first two diagonals of the matrix). Then adaptive coarsegraining of the matrices was performed to reduce noise using cooltools.numutils.adaptive_coarsegrain, with the three lowest coverage bins masked and a cutoff of ten reads. The level of coarsegraining for all ESC matrices was determined using the merged wild-type data to ensure identical bin sizes across conditions. For comparison of 5C matrices across conditions, we performed observed/expected normalisation by dividing each diagonal of the matrix by its mean. For high resolution, zoomed in (15 kb) heatmaps in Fig. 2 and Fig. S2 raw data were used, with comparison zoomed in heatmaps normalised to total read count of compared limb anterior and posterior tissue samples. All 5C heatmaps in the figures contain the summed read counts of at least two biological replicates apart from E11.5 embryos in Fig. 1, each individual replicate is shown in supplemental figures associated with the main figures. The number of total reads and of used reads is provided for each experiment in Table S10.

For insulation score analysis, we used raw 5C data without coarsening and applied cooltools.numutils.insul_diamond_dense to it with window=25 and without normalisation by median. The curves were further smoothed using LOWESS implementation from statsmodels.nonparametric.smoothers_lowess.lowess with frac=0.2, and plotted after inversion, as in raw insulation score valleys correspond to peaks of insulation, and peaks are easier to interpret visually. cooltools.lib.peaks.peakdet was used to determine location of peaks in inverted smoothed data with prominence of at least 0.2, and they were shown below the plots.

Acknowledgements

We thank the staff of the Institute of Genetics and Molecular Medicine advanced imaging resource and technical services for their assistance with imaging and sequencing. We also thank Lorraine Rose, Kyle Davies and the staff at the Biomedical Research Facility/Evans Building for expert technical assistance. We thank Maxim Imakaev for help with coarsegraining.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: I.W., R.E.H., W.A.B., L.A.L.; Methodology: I.W., L.K., P.S.D., E.A., F.K., R.E.H., L.A.L.; Validation: I.W., L.A.L.; Formal analysis: I.W., L.K., I.M.F., L.A.L.; Investigation: I.W., L.K., P.S.D., E.A., F.K., L.A.L.; Data curation: I.W., L.K., L.A.L.; Writing - original draft: I.W., L.K., L.A.L.; Writing - review & editing: I.W., L.K., I.M.F., R.E.H., W.A.B., L.A.L.; Visualization: I.W., L.K., L.A.L.; Supervision: I.W., R.E.H., W.A.B., L.A.L.; Funding acquisition: R.E.H., W.A.B.

Funding

L.K. and E.A. were funded by PhD studentships from the UK Medical Research Council (MRC). I.M.F. is funded by a PhD studentship from The Darwin Trust of Edinburgh. Work in the W.A.B. lab is funded by an MRC University Unit grant (MC_UU_00007/2) and in the R.E.H. lab by an MRC University Unit grant (MM_UU_00007/8).

Data availability

5C datasets have been deposited in GEO under accession number GSE135840.

Supplementary information

Supplementary information available online at <http://dev.biologists.org/lookup/doi/10.1242/dev.179523.supplemental>

References

- Anderson, E. and Hill, R. E. (2014). Long range regulation of the sonic hedgehog gene. *Curr. Opin. Genet. Dev.* **27**, 54–59. doi:10.1016/j.gde.2014.03.011
- Anderson, E., Devenney, P. S., Hill, R. E. and Lettice, L. A. (2014). Mapping the *Shh* long-range regulatory domain. *Development* **141**, 3934–3943. doi:10.1242/dev.108480
- Araki, K., Araki, M., Miyazaki, J. and Vassalli, P. (2006). Site-specific recombination of a transgene in fertilized eggs by transient expression of Cre recombinase. *Proc. Natl. Acad. Sci. USA* **92**, 160–164. doi:10.1073/pnas.92.1.160
- Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., Grimes, G. R., Therizols, P. and Bickmore, W. A. (2019). Decreased enhancer-promoter proximity accompanying enhancer activation. *Mol. Cell.* (in press). doi:10.1016/j.molcel.2019.07.038
- Berlivet, S., Paquette, D., Dumouchel, A., Langlais, D., Dostie, J. and Kmita, M. (2013). Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet.* **9**, e1004018. doi:10.1371/journal.pgen.1004018
- Chiang, C., Litingtung, Y., Lee, E., Young, K. E., Corden, J. L., Westphal, H. and Beachy, P. A. (1996). Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature* **383**, 407–413. doi:10.1038/383407a0
- de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. L. and de Laat, W. (2015). CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**, 676–684. doi:10.1016/j.molcel.2015.09.023
- Despang, A., Schöpfung, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Witter, L., Vingron, M. et al. (2019). Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271. doi:10.1038/s41588-019-0466-z
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380. doi:10.1038/nature11082

- Dostie, J. and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. *Nat. Protoc.* **2**, 988–1002. doi:10.1038/nprot.2007.116
- Dostie, J., Zhan, Y. and Dekker, J. (2007). Chromosome conformation capture carbon copy technology. *Curr. Protoc. Mol. Biol.* **80**, 21.14.1–21.14.13. doi:10.1002/0471142727.mb2114s80
- Eskeland, R., Leeb, M., Grimes, G. R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skultchi, A. I., Wutz, A. et al. (2010). Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38**, 452–464. doi:10.1016/j.molcel.2010.02.032
- Fabre, P. J., Leleu, M., Mormann, B. H., Lopez-Delisle, L., Noordermeer, D., Beccari, L. and Duboule, D. (2017). Large scale genomic reorganization of topological domains at the HoxD locus. *Genome Biol.* **18**, 149. doi:10.1186/s13059-017-1278-z
- Ferraiuolo, M. A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X. Q. D., Nadler, M., Blanchette, M. and Dostie, J. (2010). The three-dimensional architecture of Hox cluster silencing. *Nucleic Acids Res.* **38**, 7472–7484. doi:10.1093/nar/gkq644
- Flavahan, W. A., Drier, Y., Liao, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachaminov, A. O., Suvà, M. L. and Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114. doi:10.1038/nature16490
- Frankie, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpfli, R., Kraft, K., Kemper, R., Jerković, I., Chan, W.-L. et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269. doi:10.1038/nature19800
- Fraser, J., Rousseau, M., Blanchette, M. and Dostie, J. (2010). *Computing Chromosome Conformation*, pp. 251–268. Totowa, NJ: Humana Press.
- Fraser, J., Ethier, S. D., Miura, H. and Dostie, J. (2012). A torrent of data: Mapping chromatin organization using 5C and high-throughput sequencing. *Methods Enzymol.* **513**, 113–141. doi:10.1016/B978-0-12-391938-0.00005-7
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049. doi:10.1016/j.celrep.2016.04.085
- Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai, Y., Tang, Y. et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910. doi:10.1016/j.cell.2015.07.038
- Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., Krüger, P. H. L., Teunissen, H., Medema, R. H., van Steensel, B. et al. (2017). The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707.e14. doi:10.1016/j.cell.2017.04.013
- Hirate, Y. and Okamoto, H. (2006). Canopy1, a novel regulator of FGF signaling around the midbrain-hindbrain boundary in zebrafish. *Curr. Biol.* **16**, 421–427. doi:10.1016/j.cub.2006.01.055
- Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M. and Epstein, D. J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133**, 761–772. doi:10.1242/dev.02239
- Lajoie, B. R., van Berkum, N. L., Sanyal, A. and Dekker, J. (2009). My5C: web tools for chromosome conformation capture studies. *Nat. Methods* **6**, 690–691. doi:10.1038/nmeth1009-690
- Lee, S.-K., Jurata, L. W., Funahashi, J., Ruiz, E. C. and Pfaff, S. L. (2004). Analysis of embryonic motoneuron gene regulation: derepression of general activators function in concert with enhancer factors. *Development* **131**, 3295–3306. doi:10.1242/dev.01179
- Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735. doi:10.1093/hmg/ddg180
- Lettice, L. A., Williamson, I., Devenney, P. S., Kilanowski, F., Dorin, J. and Hill, R. E. (2014). Development of five digits is controlled by a bipartite long-range cis-regulator. *Development* **141**, 1715–1725. doi:10.1242/dev.095430
- Lettice, L. A., Devenney, P., De Angelis, C. and Hill, R. E. (2017). The conserved sonic hedgehog limb enhancer consists of discrete functional elements that regulate precise spatial expression. *Cell Rep.* **20**, 1396–1408. doi:10.1016/j.celrep.2017.07.037
- Li Song, D. and Joyner, A. L. (2000). Two Pax2/5/8-binding sites in Engrailed2 are required for proper initiation of endogenous mid-hindbrain expression. *Mech. Dev.* **90**, 155–165. doi:10.1016/S0925-4773(99)00242-7
- Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Ojitz, J. M., Laxova, R. et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025. doi:10.1016/j.cell.2015.04.004
- Morey, C., Da Silva, N. R., Perry, P. and Bickmore, W. A. (2007). Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development* **134**, 909–919. doi:10.1242/dev.02779
- Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O. and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021. doi:10.1126/science.1262088
- Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., Sedat, J. et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385. doi:10.1038/nature11049
- Nora, E. P., Goloborodko, A., Valton, A.-L., Dekker, J., Mirny, L. A., Bruneau Correspondence, B. G., Gibcus, J. H., Ueberohrn, A., Abdennur, N. and Bruneau, B. G. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–933.e22. doi:10.1016/j.cell.2017.05.004
- Paliou, C., Guckelberger, P., Schöpfli, R., Heinrich, V., Esposito, A., Chiariello, A. M., Bianco, S., Annunziata, C., Helmuth, J., Haas, S. et al. (2019). Preformed chromatin topology assists transcriptional robustness of *Shh* during limb development. *Proc. Natl. Acad. Sci. USA* **116**, 12390–12399. doi:10.1073/pnas.1900672116
- Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308. doi:10.1038/nprot.2013.143
- Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680. doi:10.1016/j.cell.2014.11.021
- Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D. et al. (2017). Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320.e24. doi:10.1016/j.cell.2017.09.026
- Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P. J., Beccari, L., El-Idrissi, I., Nguyen Huynh, T. H. N., Ozadam, H., Dekker, J. and Duboule, D. (2017). The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* **31**, 2264–2281. doi:10.1101/gad.307769.117
- Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., Wong, M. C., Maddren, M., Zhang, R., Heitner, S. G. et al. (2013). ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63. doi:10.1093/nar/gks1172
- Sagai, T., Hosoya, M., Mizushima, Y., Tamura, M. and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific *Shh* expression and truncation of the mouse limb. *Development* **132**, 797–803. doi:10.1242/dev.01613
- Sagai, T., Amaro, T., Tamura, M., Mizushima, Y., Sumiyama, K. and Shiroishi, T. (2009). A cluster of three long-range enhancers directs regional *Shh* expression in the epithelial linings. *Development* **136**, 1665–1674. doi:10.1242/dev.032714
- Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J. et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465. doi:10.1073/pnas.1518552112
- Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L. et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56. doi:10.1038/nature24281
- Soshnikova, N., Montavon, T., Leleu, M., Gajjar, N. and Duboule, D. (2010). Functional analysis of CTCF during mammalian limb development. *Dev. Cell* **19**, 819–830. doi:10.1016/j.devcel.2010.11.009
- Sun, F., Chronis, C., Kronenberg, M., Chen, X.-F., Su, T., Lay, F. D., Plath, K., Kurdiani, S. K. and Carey, M. F. (2019). Promoter-enhancer communication occurs primarily within insulated neighborhoods. *Mol. Cell* **73**, 250–263.e5. doi:10.1016/j.molcel.2018.10.039
- Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W. and Spitz, F. (2016). The *Shh* topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Dev. Cell* **39**, 529–543. doi:10.1016/j.devcel.2016.10.015
- Vian, L., Pekowska, A., Rao, S. S. P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khatabi, L., Dose, M., Pruett, N. et al. (2018). The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178.e20. doi:10.1016/j.cell.2018.03.072
- Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R. S., Paquette, D., Dostie, J. and Bickmore, W. A. (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* **28**, 2778–2791. doi:10.1101/gad.251694.114
- Williamson, I., Lettice, L. A., Hill, R. E. and Bickmore, W. A. (2016). *Shh* and ZRS enhancer colocalisation is specific to the zone of polarising activity. *Development* **143**, 2994–3001. doi:10.1242/dev.139188
- Zelenchuk, T. A. and Brusés, J. L. (2011). In Vivo labeling of zebrafish motor neurons using an *mnx1* enhancer and Gal4/UAS. *Genesis* **49**, 546–554. doi:10.1002/dvg.20766

References

- Akhtar, W., De Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., De Ridder, J., Berns, A., Wessels, L.F.A., Van Lohuizen, M., and Van Steensel, B. (2013). Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 154, 914–927.
- Alappat, S., Zhang, Z.Y., and Chen, Y.P. (2003). Msx homeobox gene family and craniofacial development. *Cell Res.* 13, 429–442.
- Albiez, H., Cremer, M., Tiberi, C., Vecchio, L., Schermelleh, L., Dittrich, S., Küpper, K., Joffe, B., Thormeyer, T., Von Hase, J., et al. (2006). Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. *Chromosom. Res.* 14, 707–733.
- Alexander, J.M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S., and Weiner, O.D. (2019). Live-cell imaging reveals enhancer-dependent sox2 transcription in the absence of enhancer proximity. *Elife* 8, 1–42.
- Alvarez-Medina, R., Cayuso, J., Okubo, T., Takada, S., and Martí, E. (2008). Wnt canonical pathway restricts graded Shh/Gli patterning activity through the regulation of Gli3 expression. *Development* 135, 237–247.
- Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* 16, 47–57.
- Anderson, E., Peluso, S., Lettice, L.A., and Hill, R.E. (2012). Human limb abnormalities caused by disruption of hedgehog signaling. *Trends Genet.* 28, 364–373.
- Anderson, E., Devenney, P.S., Hill, R.E., and Lettice, L.A. (2014). Mapping the Shh long-range regulatory domain. *Development* 707, 108480.
- Andrey, G., and Mundlos, S. (2017). The three-dimensional genome: Regulating gene expression during pluripotency and development. *Dev.* 144, 3646–3658.
- Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* (80-.). 340.
- Andrey, G., Schopflin, R., Jerkovic, I., Heinrich, V., Ibrahim, D.M., Paliou, C., Hochradel, M., Timmermann, B., Haas, S., Vingron, M., et al. (2017). Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res.* 27, 223–233.
- Arenas, E., Denham, M., and Villaescusa, J.C. (2015). How to make a midbrain dopaminergic neuron. *Dev.* 142, 1918–1936.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18, 285–298.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., De Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543, 519–524.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., James Kent, W., Mattick, J.S., and Haussler, D. (2004). Ultraconserved Elements in the Human Genome Downloaded from.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.* 109, 21301–21306.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* 405, 482–485.

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

Belloni, E., Muenke, M., Roessler, E., Traverso, G., Siegel-Bartelt, J., Frumkin, A., Mitchell, H.F., Donis-Keller, H., Helms, C., Hing, A. V., et al. (1996). Identification of sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nat. Genet.* 14, 353–356.

Benabdallah, N.S., and Bickmore, W.A. (2016). Regulatory domains and their mechanisms. *Cold Spring Harb. Symp. Quant. Biol.* 80, 45–51.

Benabdallah, N.S., Gautier, P., Hekimoglu-Balkan, B., Lettice, L.A., Bhatia, S., and Bickmore, W.A. (2016). SBE6: a novel long-range enhancer involved in driving sonic hedgehog expression in neural progenitor cells. *Open Biol.* 6, 160197.

Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D., Grimes, G.R., Therizols, P., and Bickmore, W.A. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol. Cell* 76, 473-484.e7.

Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., and Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.* 25, 2453–2464.

Berlivet, S., Paquette, D., Dumouchel, A., Langlais, D., Dostie, J., and Kmita, M. (2013). Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs. *PLoS Genet.* 9, e1004018.

- Bhatia, S., Jan Kleinjan, D., Uttley, K., Mann, A., Dellepiane, N., and Bickmore, W.A. (2020). Q-STARZ: Quantitative Spatial and Temporal Assessment of Regulatory element activity in Zebrafish. *BioRxiv* 2020.09.13.290460.
- Bialkowska, A.B., Yang, V.W., and Mallipattu, S.K. (2017). Krüppel-like factors in mammalian stem cells and development. *Dev.* 144, 737–754.
- Bickmore, W.A., and Van Steensel, B. (2013). Genome architecture: Domain organization of interphase chromosomes. *Cell* 152, 1270–1284.
- Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* (80-.). 362.
- Bitgood, M.J., Shen, L., and McMahon, A.P. (1996). Sertoli cell signaling by Desert hedgehog regulates the male germline. *Curr. Biol.* 6, 298–304.
- Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-seq identification of weakly conserved heart enhancers. *Nat. Genet.* 42, 806–812.
- Boija, A., Klein, I.A., Sabari, B.R., Dall’Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* 175, 1842-1855.e16.
- Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R., et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* 3, 0826–0842.
- Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* 171, 557-572.e24.
- Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* 132, 311–322.
- Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A., and Bickmore, W.A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* 10, 211–219.
- Boyle, S., Flyamer, I.M., Williamson, I., Sengupta, D., Bickmore, W.A., and Illingworth, R.S. (2020). A central role for canonical PRC1 in shaping the 3D nuclear landscape. *Genes Dev.* 34, 1–19.
- Briscoe, J., Chen, Y., Jessell, T.M., and Struhl, G. (2001). A hedgehog-insensitive form of Patched provides evidence for direct long-range morphogen activity of Sonic

hedgehog in the neural tube. *Mol. Cell* 7, 1279–1291.

Britto, J., Tannahill, D., and Keynes, R. (2002). A critical role for sonic hedgehog signaling in the early expansion of the developing brain. *Nat. Neurosci.* 5, 103–110.

Buckley, M.S., and Lis, J.T. (2014). Imaging RNA Polymerase II transcription sites in living cells. *Curr. Opin. Genet. Dev.* 25, 126–130.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218.

Burcin, M., Rü, † Rü'diger, Arnold, R., Lutz, M., Kaiser, B., Runge, ‡ Dieter, Lottspeich, § Friedrich, Filippova, G.N., Lobanenko, V. V., and Renkawitz, R. (1997). Negative Protein 1, Which Is Required for Function of the Chicken Lysozyme Gene Silencer in Conjunction with Hormone Receptors, Is Identical to the Multivalent Zinc Finger Repressor CTCF. *Mol. Cell. Biol.* 17, 1281–1288.

Burke, R., Nellen, D., Bellotto, M., Hafen, E., Senti, K.A., Dickson, B.J., and Basler, K. (1999). Dispatched, a novel sterol-sensing domain protein dedicated to the release of cholesterol-modified Hedgehog from signaling cells. *Cell* 99, 803–815.

Busslinger, G.A., Stocsits, R.R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.-M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* 544, 503–507.

Chambeyron, S., and Bickmore, W.A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* 18, 1119–1130.

Chan, C.S.Y., Lonfat, N., Zhao, R., Davis, A.E., Li, L., Wu, M.R., Lin, C.H., Ji, Z., Cepko, C.L., and Wang, S. (2020). Cell type- And stage-specific expression of Otx2 is regulated by multiple transcription factors and cis-regulatory modules in the retina. *Dev.* 147, 1–13.

Chen, Y., and Struhl, G. (1998). In vivo evidence that Patched and Smoothed constitute distinct binding and transducing components of a Hedgehog receptor complex. *Development* 125.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018a). Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.*

Chen, J., Wildhardt, G., Zhong, Z., Röth, R., Weiss, B., Steinberger, D., Decker, J., Blum, W.F., and Rappold, G. (2009). Enhancer deletions of the SHOX gene as a frequent cause of short stature: The essential role of a 250 kb downstream regulatory domain. *J. Med. Genet.* 46, 834–839.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., Van Steensel, B., Ma, J., and Belmont, A.S. (2018b). Mapping 3D genome

organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* 217, 4025–4048.

Chiang, C., Litingtung, Y., Lee, E., Young, K.E., Corden, J.L., Westphal, H., and Beachy, P.A. (1996). Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature* 383, 407–413.

Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* (80-.). 361, 412–415.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* (80-.). 361.

Crane-Smith, Z., Schoenebeck, J., Graham, K.A., Devenney, P.S., Rose, L., Ditzell, M., Klenin, N., Kurrasch, D.M., Lettice, L.A., and Hill, R.E. (2019). A Highly Conserved Shh Enhancer Coordinates Hypothalamic and Craniofacial Development. *BioRxiv Dev. Biol.* 794198.

Cremer, T., and Cremer, M. (2010). Chromosome territories. *Cold Spring Harb. Perspect. Biol.* 2, a003889.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21931–21936.

Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P., and Bickmore, W.A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.* 145, 1119–1131.

Cuartero, S., Weiss, F.D., Dharmalingam, G., Guo, Y., Ing-Simmons, E., Masella, S., Robles-Rebollo, I., Xiao, X., Wang, Y.F., Barozzi, I., et al. (2018). Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nat. Immunol.* 19, 932–941.

Davidson, I.F., Goetz, D., Zaczek, M.P., Molodtsov, M.I., Huis in 't Veld, P.J., Weissmann, F., Litos, G., Cisneros, D.A., Ocampo-Hafalla, M., Ladurner, R., et al. (2016). Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* 35, 2671–2685.

Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J.M. (2019). DNA loop extrusion by human cohesin. *Science* (80-.). 366, 1338–1345.

Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A., et al. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* 158, 849–860.

Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.L., Timmermann, B., Wittler, L., Vingron, M., et al. (2019). Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* 51, 1263–1271.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.

Dorigo, B., Schalch, T., Kulangara, A., Duda, S., Schroeder, R.R., and Richmond, T.J. (2004). Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* (80-.). 306, 1571–1573.

Downen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159, 374–387.

Doyle, E.L., Booher, N.J., Standage, D.S., Voytas, D.F., Brendel, V.P., Vandyk, J.K., and Bogdanove, A.J. (2012). TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res.* 40, W117-22.

Dubourg, C., Bendavid, C., Pasquier, L., Henry, C., Odent, S., and David, V. (2007). Holoprosencephaly. *Orphanet J. Rare Dis.* 2, 8.

Echelard, Y., Epstein, D.J., St-Jacques, B., Shen, L., Mohler, J., McMahon, J.A., and McMahon, A.P. (1993). Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. *Cell* 75, 1417–1430.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Epstein, D.J., McMahon, A.P., and Joyner, A.L. (1999). Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms. *Development* 126, 281–292.

Eres, I.E., and Gilad, Y. (2020). A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends Genet.* 37, 216–223.

Fabre, P.J., Leleu, M., Mormann, B.H., Lopez-Delisle, L., Noordermeer, D., Beccari, L., and Duboule, D. (2017). Large scale genomic reorganization of topological domains at the HoxD locus. *Genome Biol.* 18, 149.

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I., et al. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* 570, 395–399.

Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenko, V. V (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to

bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* 16, 2802–2813.

Finch, J.T., and Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U. S. A.* 73, 1897–1901.

Flavahan, W.A., Drier, Y., Liao, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529, 110–114.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S. V, Abdennur, N., Razin, S. V, Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nat. Publ. Gr.*

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538, 265–269.

Freire-Maia, A., Freire-Maia, N., Morton, E, N., Azevedo, S, E., and Quelce-Salgado, A. (1975). Genetics of acheiropodia (“the handless and footless families of Brazil”). *Am. J. Hum. Genet.* 27, 521–527.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* 15, 2038–2049.

Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., and Mirny, L.A. (2017). Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb. Symp. Quant. Biol.* 82, 45–55.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell* 166, 358–368.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. *Science* 361, 1341–1345.

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science* (80-.). 360, 102–105.

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korb, J.O., and Furlong, E.E.M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* 1.

De Gobbi, M., Anguita, E., Hughes, J., Sloane-Stanley, J.A., Sharpe, J.A., Koch, C.M., Dunham, I., Gibbons, R.J., Wood, W.G., and Higgs, D.R. (2007). Tissue-specific histone modification and transcription factor binding in α globin gene expression. *Blood* 110, 4503–4510.

Goode, D.K., Snell, P., Smith, S.F., Cooke, J.E., and Elgar, G. (2005). Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* 86, 172–181.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., De Klein, A., Wessels, L., De Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948–951.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* 162, 900–910.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* 169, 693–707.e14.

Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* 6.

Hanssen, L.L.P.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M., Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R., et al. (2017). Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* 19, 952–961.

Hardy, R.Y., and Resh, M.D. (2012). Identification of N-terminal residues of sonic hedgehog important for palmitoylation by Hedgehog acyltransferase. *J. Biol. Chem.* 287, 42881–42889.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405, 486–489.

He, H.H., Meyer, C.A., Shin, H., Bailey, S.T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., et al. (2010). Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* 42, 343–347.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., and Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17507–17512.

Hehr, U., Gross, C., Diebold, U., Wahl, D., Beudt, U., Heidemann, P., Hehr, A., and Mueller, D. (2004). Wide phenotypic variability in families with holoprosencephaly and a sonic hedgehog mutation. *Eur. J. Pediatr.* 163, 347–352.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive

chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* 38, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* 16, 144–154.

Henikoff, S., and Smith, M.M. (2015). Histone variants and epigenetics. *Cold Spring Harb. Perspect. Biol.* 7.

Hill, R.E. (2007). How to make a zone of polarizing activity: Insights into limb development via the abnormality preaxial polydactyly. *Dev. Growth Differ.* 49, 439–448.

Hirate, Y., and Okamoto, H. (2006). Canopy1, a novel regulator of FGF signaling around the midbrain-hindbrain boundary in zebrafish. *Curr. Biol.* 16, 421–427.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. *Cell* 155.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* (80-.). 351, 1454–1458.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). Perspective A Phase Separation Model for Transcriptional Control.

Hughes, J.R., Cheng, J.F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., De Jong, P., Rubin, E., and Higgs, D.R. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl. Acad. Sci. U. S. A.* 102, 9830–9835.

Hui, C.C., Slusarski, D., Platt, K.A., Holmgren, R., and Joyner, A.L. (1994). Expression of three mouse homologs of the drosophila segment polarity gene cubitus interruptus, Gli, Gli-2, and Gli-3, in ectoderm- and mesoderm-derived tissues suggests multiple roles during postimplantation development. *Dev. Biol.* 162, 402–413.

Ianakev, P., Van Baren, M.J., Daly, M.J., Toledo, S.P.A., Cavalcanti, M.G., Neto, J.C., Silveira, E.L., Freire-Maia, A., Heutink, P., Kilpatrick, M.W., et al. (2001). Acheiropodia is caused by a genomic deletion in C7orf2, the human orthologue of the

Lmbr1 gene. *Am. J. Hum. Genet.* 68, 38–45.

Inestrosa, N.C., and Varela-Nallar, L. (2015). Wnt signalling in neuronal differentiation and development. *Cell Tissue Res.* 359, 215–223.

Islam, M.E., Kikuta, H., Inoue, F., Kanai, M., Kawakami, A., Parvin, M.S., Takeda, H., and Yamasu, K. (2006). Three enhancer regions regulate gbx2 gene expression in the isthmus region during zebrafish development. *Mech. Dev.* 123, 907–924.

Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M., and Epstein, D.J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* 133, 761–772.

Jeong, Y., Leskow, F.C., El-Jaick, K., Roessler, E., Muenke, M., Yocum, A., Dubourg, C., Li, X., Geng, X., Oliver, G., et al. (2008a). Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat. Genet.* 40, 1348–1353.

Jeong, Y., Leskow, F.C., El-Jaick, K., Roessler, E., Muenke, M., Yocum, A., Dubourg, C., Li, X., Geng, X., Oliver, G., et al. (2008b). Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat. Genet.* 40, 1348–1353.

Jeong, Y., Dolson, D.K., Waclaw, R.R., Matise, M.P., Sussel, L., Campbell, K., Kaestner, K.H., and Epstein, D.J. (2011). Spatial and temporal requirements for sonic hedgehog in the regulation of thalamic interneuron identity. *Development* 138, 531–541.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61.

Jubb, A.W., Young, R.S., Hume, D.A., and Bickmore, W.A. (2016). Enhancer Turnover Is Associated with a Divergent Transcriptional Response to Glucocorticoid in Mouse and Human Macrophages. *J. Immunol.* 196, 813–822.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* 10, 1–10.

Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J., and Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. *Science* (80-.). 366, 1345–1349.

Kioussis, D., Vanin, E., Delange, T., Flavell, R.A., and Grosveld, F.G. (1983). β -Globin gene inactivation by DNA translocation in $\gamma\beta$ -thalassaemia. *Nature* 306, 662–666.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenko, V. V (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol. Cell. Biol.* 13, 7612–7624.

- Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell* 128, 693–705.
- Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schöpflin, R., Markowski, J., Ibrahim, D.M., Acuna-Hidalgo, R., Despag, A., Andrey, G., et al. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* 21, 305–310.
- Krauss, S., Concordet, J.P., and Ingham, P.W. (1993). A functionally conserved homolog of the *Drosophila* segment polarity gene *hh* is expressed in tissues with polarizing activity in zebrafish embryos. *Cell* 75, 1431–1444.
- Krivega, I., and Dean, A. (2017). LDB1-mediated enhancer looping can be established independent of mediator and cohesin. *Nucleic Acids Res.* 45, 8255–8268.
- Kruger, L.M., and Kumar, A. (1994). Acheiropody. A report of two cases. *J. Bone Jt. Surg.* 76, 1557–1560.
- Krüger, M., Mennerich, D., Fees, S., Schäfer, R., Mundlos, S., and Braun, T. (2001). Sonic hedgehog is a survival factor for hypaxial muscles during mouse development. *Development* 128, 743–752.
- Kruithof, M., Chien, F.T., Routh, A., Logie, C., Rhodes, D., and Van Noort, J. (2009). Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nat. Struct. Mol. Biol.* 16, 534–540.
- Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Ye, Z., Li, B., Dixon, J.R., Desai, A., et al. (2017). Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells 2 3.
- Kurokawa, D., Takasaki, N., Kiyonari, H., Nakayama, R., Kimura-Yoshida, C., Matsuo, I., and Aizawa, S. (2004a). Regulation of *Otx2* expression and its functions in mouse epiblast and anterior neuroectoderm. *Development* 131, 3307–3317.
- Kurokawa, D., Kiyonari, H., Nakayama, R., Kimura-Yoshida, C., Matsuo, I., and Aizawa, S. (2004b). Regulation of *Otx2* expression and its functions in mouse forebrain and midbrain. *Development* 131, 3319–3331.
- Kurokawa, D., Ohmura, T., Sakurai, Y., Inoue, K., Suda, Y., and Aizawa, S. (2014). *Otx2* expression in anterior neuroectoderm and forebrain/midbrain is directed by more than six enhancers. *Dev. Biol.* 387, 203–213.
- Kuwahara, A., Sakai, H., Xu, Y., Itoh, Y., Hirabayashi, Y., and Gotoh, Y. (2014). Tcf3 represses Wnt- β -catenin signaling and maintains neural stem cell population during neocortical development. *PLoS One* 9.
- Labialle, S., Marty, V., Bortolin-Cavaillé, M.-L., Hoareau-Osman, M., Pradère, J.-P., Valet, P., Martin, P.G.P., and Cavaillé, J. (2014). The miR- 379 / miR- 410 cluster at the imprinted *Dlk 1 -Dio 3* domain controls neonatal metabolic adaptation. *EMBO J.* 33, 2216–2230.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1 α suggests a role for phase separation in heterochromatin. *Nature* 547, 236–240.

Lee, J.J., Ekker, S.C., Von Kessler, D.P., Porter, J.A., Sun, B.I., and Beachy, P.A. (1994). Autoproteolysis in hedgehog protein biogenesis. *Science* (80-.). 266, 1528–1537.

Leth Bak, A., Zeuthen, J., and Crick, F.H.C. (1977). Higher-order structure of human mitotic chromosomes.

Lettice, L. a, Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B. a, Endo, N., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7548–7553.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12, 1725–1735.

Lettice, L.A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P.S., Gautier, P., Morrison, H., Fantes, J., Hill, R.E., and Fitzpatrick, D.R. (2011). Enhancer-adoption as a mechanism of human developmental disease. *Hum. Mutat.* 32, 1492–1499.

Lettice, L.A., Williamson, I., Wiltshire, J.H., Peluso, S., Devenney, P.S., Hill, A.E., Essafi, A., Hagman, J., Mort, R., Grimes, G., et al. (2012). Opposing Functions of the ETS Factor Family Define Shh Spatial Expression in Limb Buds and Underlie Polydactyly. *Dev. Cell* 22, 459–467.

Lettice, L.A., Williamson, I., Devenney, P.S., Kilanowski, F., Dorin, J., and Hill, R.E. (2014). Development of five digits is controlled by a bipartite long-range cis-regulator. *Development* 141, 1715–1725.

Lettice, L.A., Devenney, P., De Angelis, C., and Hill, R.E. (2017). The Conserved Sonic Hedgehog Limb Enhancer Consists of Discrete Functional Elements that Regulate Precise Spatial Expression. *Cell Rep.* 20, 1396–1408.

Li, Y., Haarhuis, J.H.I., Sedeño Cacciatore, Á., Oldenkamp, R., van Ruiten, M.S., Willems, L., Teunissen, H., Muir, K.W., de Wit, E., Rowland, B.D., et al. (2020). The structural basis for cohesin–CTCF-anchored loops. *Nature* 578, 472–476.

Li Song, D., and Joyner, A.L. (2000). Two Pax2/5/8-binding sites in Engrailed2 are required for proper initiation of endogenous mid-hindbrain expression. *Mech. Dev.* 90, 155–165.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009).

Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* (80-.). 326, 289–293.

Lim, B., Heist, T., Levine, M., and Fukaya, T. (2018). Visualization of Transvection in Living *Drosophila* Embryos. *Mol. Cell* 70, 287-296.e6.

Litingtung, Y., Lei, L., Westphal, H., and Chiang, C. (1998). Sonic hedgehog is essential to foregut development. *Nat. Genet.* 20, 58–61.

Liu, N.Q., Maresca, M., van den Brand, T., Braccioli, L., Schijns, M.M.G.A., Teunissen, H., Bruneau, B.G., Nora, E.P., and de Wit, E. (2020). WAPL maintains a cohesin loading cycle to preserve cell-type-specific distal gene regulation. *Nat. Genet.*

Lobanenko, V. V., Nicolas, R.H., Adler, V. V., Paterson, H., Klenova, E.M., Polotskaja, A. V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* 5, 1743–1753.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* 167, 1170–1187.

Long, H.K., Osterwalder, M., Welsh, I.C., Hansen, K., Davies, J.O.J., Liu, Y.E., Koska, M., Adams, A.T., Aho, R., Arora, N., et al. (2020). Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell* 27, 765-783.e14.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260.

Lupiañez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025.

Marigo, V., Davey, R.A., Zuo, Y., Cunningham, J.M., and Tabin, C.J. (1996). Biochemical evidence that patched is the hedgehog receptor. *Nature* 384, 176–179.

Maroulakou, I.G., and Bowe, D.B. (2000). Expression and function of Ets transcription factors in mammalian development: A regulatory network. *Oncogene* 19, 6432–6442.

Mason, S., Piper, M., Gronostajski, R.M., and Richards, L.J. (2009). Nuclear factor one transcription factors in CNS development. *Mol. Neurobiol.* 39, 10–23.

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nat.* 2020 5837818 583, 699–710.

Moore, J.M., Rabaia, N.A., Smith, L.E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C.M., Collins, S.J., Kemp, C.J., Lobanenko, V. V., et al. (2012). Loss of

maternal CTCF is associated with peri-implantation lethality of CtCf null embryos. *PLoS One* 7, e34915.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M.P., and Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* 9, 6047–6068.

Morey, C., Da Silva, N.R., Perry, P., and Bickmore, W.A. (2007). Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. *Development* 134, 909–919.

Morey, C., Kress, C., and Bickmore, W.A. (2009). Lack of bystander activation shows that localization exterior to chromosome territories is not sufficient to up-regulate gene expression. *Genome Res.* 19, 1184–1194.

Murayama, Y., and Uhlmann, F. (2014). Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature* 505, 367–371.

Murayama, Y., and Uhlmann, F. (2015). DNA Entry into and Exit out of the Cohesin Ring by an Interlocking Gate Mechanism. *Cell* 163, 1628–1640.

Murdoch, J.N., and Copp, A.J. (2010). The relationship between sonic Hedgehog signaling, cilia, and neural tube defects. *Birth Defects Res. Part A Clin. Mol. Teratol.* 88, 633–652.

Murone, M., Rosenthal, A., and De Sauvage, F.J. (1999). Sonic hedgehog signaling by the patched-smoothened receptor complex. *Curr. Biol.* 9, 76–84.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* (80-.). 347, 1017–1021.

Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Leading Edge Review Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes.

Nasmyth, K. (2001). Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. *Annu. Rev. Genet.* 35, 673–745.

Natsume, T., Kiyomitsu, T., Saga, Y., and Kanemaki, M.T. (2016). Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep.* 15, 210–218.

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., and Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat. Methods* 6, 917–922.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning Human Gene Deserts for Long-Range Enhancers. *Science* (80-.). 302, 413.

Noonan, J.P., and McCallion, A.S. (2010). Genomics of Long-Range Regulatory Elements. *Annu. Rev. Genomics Hum. Genet.* 11, 1–23.

Noordermeer, D., Branco, M.R., Splinter, E., Klous, P., Van Ijcken, W., Swagemakers, S., Koutsourakis, M., Van Der Spek, P., Pombo, A., and De Laat, W. (2008). Transcription and chromatin organization of a housekeeping gene cluster containing an integrated β -globin locus control region. *PLoS Genet.* 4, e1000016.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.-L., Dekker, J., Mirny, L.A., Bruneau Correspondence, B.G., Gibcus, J.H., Uebersohn, A., Abdennur, N., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* 169, 930-933.e22.

Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* 511, 428–434.

Nozaki, T., Imai, R., Tanbo, M., Nagashima, R., Tamura, S., Tani, T., Joti, Y., Tomita, M., Hibino, K., Kanemaki, M.T., et al. (2017). Dynamic Organization of Chromatin Domains Revealed by Super-Resolution Live-Cell Imaging. *Mol. Cell* 67, 282-293.e7.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L.A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6697–E6706.

Nüsslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in drosophila. *Nature* 287, 795–801.

Ohtsuka, T., Shimojo, H., Matsunaga, M., Watanabe, N., Kometani, K., Minato, N., and Kageyama, R. (2011). Gene Expression Profiling of Neural Stem Cells and Identification of Regulators of Neural Differentiation During Cortical Development. *Stem Cells* 29, 1817–1828.

Ong, C., and Corces, V.G. (2014). CTCF : an architectural protein bridging genome topology and function. *Nat. Publ. Gr.* 15, 234–246.

Osterwalder, M., Barozzi, I., Tissi res, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* 554, 239–243.

Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O’Shea, C.C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* (80-.). 357.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, 8-227

L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.* 15, 137–145.

Paliou, C., Guckelberger, P., Schöpflin, R., Heinrich, V., Esposito, A., Chiariello, A.M., Bianco, S., Annunziatella, C., Helmuth, J., Haas, S., et al. (2019). Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc. Natl. Acad. Sci. U. S. A.* 116, 12390–12399.

Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and De Laat, W. (2003). The β -globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.* 35, 190–194.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* 132, 422–433.

Peluso, S., Douglas, A., Hill, A., de Angelis, C., Moore, B.L., Grimes, G., Petrovich, G., Essafi, A., and Hill, R.E. (2017). Fibroblast growth factors (FGFs) prime the limb specific shh enhancer for chromatin changes that balance histone acetylation mediated by e26 transformation-specific (ETS) factors. *Elife* 6.

Pepinsky, R.B., Zeng, C., Went, D., Rayhorn, P., Baker, D.P., Williams, K.P., Bixler, S.A., Ambrose, C.M., Garber, E.A., Miatkowski, K., et al. (1998). Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* 273, 14037–14045.

Pereira, M.C.F., Brackley, C.A., Michieletto, D., Annunziatella, C., Bianco, S., Chiariello, A.M., Nicodemi, M., and Marenduzzo, D. (2018). Complementary chromosome folding by transcription factors and cohesin. *BioRxiv* 305359.

Petit, F., Sears, K.E., and Ahituv, N. (2017). Limb development: a paradigm of gene regulation. *Nat. Rev. Genet.* 18, 245–258.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. *Cell* 137, 1194–1211.

Porter, J.A., von Kessler, D.P., Ekker, S.C., Young, K.E., Lee, J.J., Moses, K., and Beachy, P.A. (1995). The product of hedgehog autoproteolytic cleavage active in local and long-range signalling. *Nature* 374, 363–366.

Porter, J.A., Ekker, S.C., Park, W.J., von Kessler, D.P., Young, K.E., Chen, C.H., Ma, Y., Woods, A.S., Cotter, R.J., Koonin, E. V, et al. (1996a). Hedgehog patterning activity: role of a lipophilic modification mediated by the carboxy-terminal autoprocessing domain. *Cell* 86, 21–34.

Porter, J.A., Young, K.E., and Beachy, P.A. (1996b). Cholesterol modification of hedgehog signaling proteins in animal development. *Science* (80-.). 274, 255–259.

Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C.A., Schneider, R., and Bickmore, W.A. (2016). Histone H3 globular domain acetylation identifies a new

class of enhancers. *Nat. Genet.* 48, 681–686.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* 174, 744–757.e24.

Racko, D., Benedetti, F., Dorier, J., and Stasiak, A. (2018). Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Res.* 46, 1648–1660.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470, 279–285.

Ragoczy, T., Bender, M.A., Telling, A., Byron, R., and Groudine, M. (2006). The locus control region is required for association of the murine β -globin locus with engaged transcription factories during erythroid maturation. *Genes Dev.* 20, 1447–1457.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell* 171, 305–320.e24.

Reddy, K.L., Zullo, J.M., Bertolino, E., and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* 452, 243–247.

Rhodes, J., Mazza, D., Nasmyth, K., and Uphoff, S. (2017). Scc2/Nipbl hops between chromosomal cohesin rings after loading. *Elife* 6.

Rhodes, J.D.P., Feldmann, A., Hernández-Rodríguez, B., Díaz, N., Brown, J.M., Fursova, N.A., Blackledge, N.P., Prathapan, P., Dobrinic, P., Huseyin, M.K., et al. (2020). Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells. *Cell Rep.* 30, 820–835.e10.

Ricci, M.A., Manzo, C., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell* 160, 1145–1158.

Riddle, R.D., Johnson, R.L., Laufer, E., and Tabin, C. (1993). Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* 75, 1401–1416.

Robinson, P.J.J., Fairall, L., Huynh, V.A.T., and Rhodes, D. (2006). EM measurements define the dimensions of the “30-nm” chromatin fiber: Evidence for a
8-229

compact, interdigitated structure. *Proc. Natl. Acad. Sci. U. S. A.* *103*, 6506–6511.

Roccio, M., Schmitter, D., Knobloch, M., Okawa, Y., Sage, D., and Lutolf, M.P. (2013). Predicting stem cell fate changes by differential cell cycle progression patterns. *Dev.* *140*, 459–470.

Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P.J., Beccari, L., El-Idrissi, I., Nguyen Huynh, T.H., Ozadam, H., Dekker, J., and Duboule, D. (2017). The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* *31*, 2264–2281.

Roessler, E., Belloni, E., Gaudenz, K., Jay, P., Berta, P., Scherer, S.W., Tsui, L.C., and Muenke, M. (1996). Mutations in the human sonic hedgehog gene cause holoprosencephaly. *Nat. Genet.* *14*, 357–360.

Ruiz i Altaba, A. (1997). Catching a Gli-mpse of hedgehog. *Cell* *90*, 193–196.

Sabari, B.R., Dall’Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* (80-.). *361*, eaar3958.

Sagai, T., Masuya, H., Tamura, M., Shimizu, K., Yada, Y., Wakana, S., Gondo, Y., Noda, T., and Shiroishi, T. (2004). Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh). *Mamm. Genome* *15*, 23–34.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., Shiroishi, T., and Balling, R. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* *132*, 797–803.

Sagai, T., Amano, T., Tamura, M., Mizushina, Y., Sumiyama, K., and Shiroishi, T. (2009). A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. *Development* *136*, 1665–1674.

Sagai, T., Amano, T., Maeno, A., Ajima, R., and Shiroishi, T. (2019). SHH signaling mediated by a prechordal and brain enhancer controls forebrain organization. *Proc. Natl. Acad. Sci. U. S. A.* *116*, 23636–23642.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* *112*, E6456–E6465.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* *489*, 109–113.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in

erythroid cells. *Nat. Genet.* 42, 53–61.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 551, 51–56.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.* 23, 2066–2077.

Seo, H., Amano, T., Seki, R., Sagai, T., Kim, J., Cho, S.W., and Shiroishi, T. (2018). Upstream Enhancer Elements of *Shh* Regulate Oral and Dental Patterning. *J. Dent. Res.* 002203451875864.

Shamseldin, H.E., Anazi, S., Wakil, S.M., Faqeih, E., El Khashab, H.Y., Salih, M.A., Al-Qattan, M.M., Hashem, M., Alsedairy, H., and Alkuraya, F.S. (2016). Novel copy number variants and major limb reduction malformation: Report of three cases. *Am. J. Med. Genet. Part A* 170, 1245–1250.

Shopland, L.S., Lynch, C.R., Peterson, K.A., Thornton, K., Kepper, N., Von Hase, J., Stein, S., Vincent, S., Molloy, K.R., Kreth, G., et al. (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J. Cell Biol.* 174, 27–38.

Shrimali, S., Srivastava, S., Varma, G., Grinberg, A., Pfeifer, K., and Srivastava, M. (2012). An ectopic CTCF-dependent transcriptional insulator influences the choice of V β gene segments for VDJ recombination at TCR β locus. *Nucleic Acids Res.* 40, 7753–7765.

Simeone, A., Avantaggiato, V., Moroni, M.C., Mavilio, F., Arra, C., Cotelli, F., Nigro, V., and Acampora, D. (1995). Retinoic acid induces stage-specific antero-posterior transformation of rostral central nervous system. *Mech. Dev.* 51, 83–98.

Simeone, A., Martinez-Barbera, J.P., Puellas, E., and Acampora, D. (2003). Role of Otx transcription factors in brain development. *Adv. Dev. Biol. Biochem.* 13, 207–250.

Singh, V.P., and Gerton, J.L. (2015). Cohesin and human disease: Lessons from mouse models. *Curr. Opin. Cell Biol.* 37, 9–17.

Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Publ. Gr.* 13.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.* 32, 3119–3129.

- Song, F., Chen, P., Sun, D., Wang, M., Dong, L., Liang, D., Xu, R.M., Zhu, P., and Li, G. (2014). Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* (80-.). 344, 376–380.
- Soshnikova, N., Montavon, T., Leleu, M., Galjart, N., and Duboule, D. (2010). Functional Analysis of CTCF During Mammalian Limb Development. *Dev. Cell* 19, 819–830.
- Van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* 18, 424–428.
- Strom, A.R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. *Nature* 547, 241–245.
- Sun, F., Chronis, C., Kronenberg, M., Chen, X.F., Su, T., Lay, F.D., Plath, K., Kurdistani, S.K., and Carey, M.F. (2019). Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. *Mol. Cell* 73, 250-263.e5.
- Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 24, 390–400.
- Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., and Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* 39, 529–543.
- Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* (80-.). 346, 1238–1242.
- Tickle, C., and Towers, M. (2017). Sonic Hedgehog Signaling in Limb Development. *Front. Cell Dev. Biol.* 5, 14.
- Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and De Laat, W. (2002). Looping and interaction between hypersensitive sites in the active β -globin locus. *Mol. Cell* 10, 1453–1465.
- Tsukiji, N., Amano, T., and Shiroishi, T. (2014). A novel regulatory element for Shh expression in the lung and gut of mouse embryos. *Mech. Dev.* 131, 127–136.
- Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.R., Jung, S., Baranello, L., Huang, S.C., El Khattabi, L., Dose, M., Pruett, N., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* 173, 1165-1178.e20.
- Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854–858.

- Vortkamp, A., Lee, K., Lanske, B., Segre, G. V., Kronenberg, H.M., and Tabin, C.J. (1996). Regulation of rate of cartilage differentiation by Indian Hedgehog and PTH-related protein. *Science* (80-.). 273, 613–622.
- Wan, L. Ben, Pan, H., Hannenhalli, S., Cheng, Y., Ma, J., Fedoriw, A., Lobanenkov, V., Latham, K.E., Schultz, R.M., and Bartolomei, M.S. (2008). Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development* 135, 2729–2738.
- Wang, S., Su, J.H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C.T., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. *Science* (80-.). 353, 598–602.
- Wassarman, K.M., Lewandoski, M., Campbell, K., Joyner, A.L., Rubenstein, J.L.R., Martinez, S., and Martin, G.R. (1997). Specification of the anterior hindbrain and establishment of a normal mid/hindbrain organizer is dependent on Gbx2 gene function. *Development* 124, 2923–2934.
- Weintraub, A.S., Li, C.H., Zamudio, A. V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171, 1573-1588.e28.
- Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.
- Widom, J., and Klug, A. (1985). Structure of the 3000Å chromatin filament: X-ray diffraction from oriented samples. *Cell* 43, 207–213.
- Wieczorek, D., Pawlik, B., Li, Y., Akarsu, N.A., Caliebe, A., May, K.J.W., Schweiger, B., Vargas, F.R., Balci, S., Gillessen-Kaesbach, G., et al. (2010). A specific mutation in the distant sonic hedgehog (SHH) cis-regulator (ZRS) causes Werner Mesomelic Syndrome (WMS) while complete ZRS duplications underlie Haas type polysyndactyly and preaxial polydactyly (PPD) with or without triphalangeal thumb. *Hum. Mutat.* 31, 81–89.
- Will, A.J., Cova, G., Osterwalder, M., Chan, W.-L., Wittler, L., Brieske, N., Heinrich, V., de Villartay, J.-P., Vingron, M., Klopocki, E., et al. (2017). Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat. Publ. Gr.* 49.
- Williamson, I., Lettice, L.A., Hill, R.E., and Bickmore, W.A. (2016). Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. *Development* 143, 2994–3001.
- Williamson, I., Kane, L., Devenney, P.S., Flyamer, I.M., Anderson, E., Kilanowski, F., Hill, R.E., Bickmore, W.A., and Lettice, L.A. (2019). Developmentally regulated Shh expression is robust to TAD perturbations. *Dev.* 146.

- Wit, E. De (2020). TADs as the Caller Calls Them. *J. Mol. Biol.* 432, 638–642.
- de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* 60, 676–684.
- Woodcock, C.L., Skoultschi, A.I., and Fan, Y. (2006). Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosom. Res.* 14, 17–25.
- Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3.
- Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J., et al. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* 36, 3573–3599.
- Yao, Y., Minor, P.J., Zhao, Y.T., Jeong, Y., Pani, A.M., King, A.N., Symmons, O., Gan, L., Cardoso, W. V., Spitz, F., et al. (2016). Cis-regulatory architecture of a brain signaling center predates the origin of chordates. *Nat. Genet.* 48, 575–580.
- Zelenchuk, T.A., and Brusés, J.L. (2011). In Vivo labeling of zebrafish motor neurons using an *mnx1* enhancer and Gal4/UAS. *Genesis* 49, 546–554.
- Zhang, D., Zeldin, D.C., and Blackshear, P.J. (2007). Regulatory factor X4 variant 3: A transcription factor involved in brain development and disease. *J. Neurosci. Res.* 85, 3515–3522.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoutte, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9.
- Zhu, J., and Mackem, S. (2011). Analysis of mutants with altered *shh* activity and posterior digit loss supports a biphasic model for *shh* function as a morphogen and mitogen. *Dev. Dyn.* 240, 1303–1310.
- Zhu, F., Nair, R.R., Fisher, E.M.C., and Cunningham, T.J. (2019). Humanising the mouse genome piece by piece. *Nat. Commun.* 10, 1–13.
- Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* 111, 996–1001.