

Deconstructing Somites across the Cervical-Thoracic Boundary using Advanced Genomics Approaches

Thesis for a PhD degree

Submitted to University of East Anglia

by Shannon A. Weldon

Principal Investigator: Prof. Andrea Münsterberg

Submission Date: 15.09.2022

DECLARATION OF OWN WORK.

I, Shannon Weldon, confirm that the work for the report with the title:

“Deconstructing somites across the cervical-thoracic boundary using advanced genomics approaches”

was undertaken by myself and that no help was provided from other sources than those allowed. All sections of the report that use quotes or describe an argument or development investigated by other scientists have been referenced, including all secondary literature used, to show that this material has been adopted to support my report.

Signature: *Shannon A. Weldon*

Date: 15/09/2022

ABSTRACT

Embryonic somites are paired segments that form on either side of the neural tube along the head-to-tail axis in vertebrates. They are of fundamental importance and generate cells of the musculoskeletal system, including myocytes, chondrocytes and tenocytes, as well as endothelial and dermal cells. In this thesis, I investigate the molecular profiles of somites with different axial identities by sequencing somites isolated from the chick embryo. I determined the accessible chromatin and transcriptome signatures of somites using ATACseq and RNAseq. In particular, I focus on axial positions that produce prospective cervical and thoracic somites, with anatomically discrete features such as ribs. In depth analysis of these data aims to uncover potential *cis*-regulatory elements and associated gene expression changes that correlate with positional identity along the cervical-thoracic boundary, as well as musculoskeletal cell fate acquisition. Classic experimental embryology combined with genomics analyses is used to validate and characterise putative enhancer elements. This provides insights into the intrinsic mechanisms that underlie morphological evolution of the vertebrate body plan.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

ACKNOWLEDGEMENTS

I am incredibly grateful for the opportunity I had to work in the Münsterberg lab for my PhD research. I would very much like to thank Andrea Münsterberg for offering me this great position in her lab. I especially want to thank her for her incredible guidance and support she has given me the past four years. I also would like to say thank you for the numerous opportunities she gave me to grow as a scientist.

I also would like to say thank you to Wilfried Hearty, Iain Macaulay, and Grant Wheeler, who supervised the progression of my PhD and were always open for questions. Furthermore, a big thanks goes to Eirini Maniou, Timothy Grocott and Ashleigh Lister for discussions and technical advice during my experiments, and more thanks to Leighton Folkes, without whom I would have no ATAC bioinformatics to analyse.

Furthermore, I would like to thank Geoffrey Mok, for not only helping me get started in the lab by teaching me relevant techniques, but also providing me with guidance and opportunities to take part in exciting experiments throughout the PhD. It was an honour to be a part of 'team somite' and I will forever be grateful for the inspiration, laughs and team spirit you gave me.

A special thanks goes to the various members of the Münsterberg, Wheeler and Grocott lab for discussion support and making my time in the UK joyful. In particular, I would like to thank Emily Smith, who was a constant and invaluable support in the lab and who was always ready to chat with me for hours over a rapidly cooling cup of tea in our kitchen about anything and everything that was on our minds.

Last, but not least, I would like to thank my parents, Stephen and Kathleen. I want to thank them for the unconditional support when I told them I was suddenly flying to England for a PhD interview that they didn't even know I was applying for, for always being there whenever I needed you, and for always pushing me and inspiring me to do whatever I want in life, even if you didn't quite understand what I was talking about. To my sister, Ashling - you have been an inspiration and a source of strength for me throughout my entire life and I want to thank you especially for being here for me. I am very lucky to have you.

Table of Contents

DECLARATION OF OWN WORK.	2
ABSTRACT	3
ACKNOWLEDGEMENTS.....	4
CHAPTER 1: INTRODUCTION	10
1.1. <i>The chick as a model organism</i>	10
1.2. <i>Development of the vertebral body plan</i>	10
1.2.1. Gastrulation	10
1.2.2. Somitogenesis.....	11
1.2.3. Clock and wavefront model.....	12
1.2.4. Somite Compartmentalisation	14
1.3. <i>The Vertebral Axial Skeleton:</i>	17
1.3.1. Differences in Axial Formula Between Species.	19
1.4. <i>Hox genes</i>	22
1.4.1. Discovery of Hox Genes.....	26
1.4.2. Vertebrate Hox genes	27
1.4.3. Interaction of Hox with lineage specifying genes	30
1.4.4. Genetic Evidence for Hox gene Function in Visceral Organs and Limbs	31
1.4.5. Homeotic transformations of the axial skeleton.	33
1.4.6. Hoxc6, Hoxc5, and Hoxc8.	35
1.4.7. Hox gene regulation.	36
1.5. <i>Enhancers</i>	37
1.6. <i>Project Aims and Objectives</i>	40
1.6.1. Aims	40
1.6.2. Objectives.....	41
CHAPTER 2: MATERIALS AND METHODS	42
2.1. <i>Embryo development and preparation</i>	42
2.1.1. Culture of chicken eggs	42
2.1.2. Embryo harvesting.....	42
2.1.2.1. For in situ hybridisation	42
2.1.2.2. For ex ovo injection and electroporation	42
2.1.2.3. For RNA extraction	42
2.1.3. Embryo dehydration	42
2.1.4. Embryo rehydration.....	43
2.2. <i>RNA probe synthesis by PCR</i>	43
2.2.1. RNA extraction	43
2.2.2. cDNA synthesis	43
2.2.3. DNA template preparation	44
2.2.3.1. Amplification of cDNA.....	44
2.2.3.2. pGEM-T-Easy cloning.....	45
2.2.3.3. Plasmid transformation.....	46
2.2.3.4. Colony PCR.....	46
2.2.4. Probe synthesis	47
2.2.5. Probe purification	48

2.3. Whole mount <i>in situ</i> hybridisation (WISH)	48
2.3.1. Day 1: Hybridisation	48
2.3.2. Day 2: Immunodetection	49
2.3.3. Day 3: Colour development	49
2.4. Histological sectioning	49
2.5. <i>In ovo</i> grafting	50
2.5.1. Part 1: donor embryo.....	50
2.5.2. Part 2: host embryo	51
2.6. RNAseq	52
2.6.1. Somite dissections	52
2.6.2. Library preparation.....	54
2.6.3. Transcriptome analysis	56
2.7. ATACseq.....	57
2.7.1. Somite dissections	57
2.7.2. Library preparation.....	57
2.7.3. Transcriptome analysis	58
2.7.4. Footprinting analysis	59
2.9. Preparation of putative enhancers	59
2.9.1. gDNA isolation.....	59
2.9.2. Plasmid constructs.....	59
2.9.3. Ex ovo plasmid injection and electroporation	61
2.9.4. <i>In ovo</i> plasmid injection and electroporation	62
2.9.5 Enhancer mutagenesis	62
2.10. <i>HOXC6</i> and <i>HOXC8</i> target validation.....	64
2.10.1. Identifying downstream targets of Hox gene.....	64
2.10.2. CRISPR-ON to activate promoters	64
2.10.2.1. Primer design	64
2.10.2.2. Plasmid construct	65
2.10.2.3. Plasmid injection and electroporation	66
CHAPTER 3: RNA-SEQUENCING ACROSS THE C-T BOUNDARY	67
3.1. RNA-Sequencing	67
3.2. Transcriptional Profiling of Cervical and Thoracic Somites.....	69
3.2.1. Quality Control Procedures	70
3.2.2. Read Alignment and Quantification	70
3.2.3. Early Analyses.....	71
3.2.4. Gene Ontology.....	72
3.2.5. Hierarchical Clustering	72
3.3. Results and Discussion	73
3.3.1 PCA.....	73
3.3.2 Differential expression.....	76
3.3.3 Heat maps	78
3.3.4 Gene Ontology.....	79
3.3.4 Additional Analyses	79
3.4. Conclusion.....	81
CHAPTER 4: IDENTIFICATION AND VALIDATION OF HOX CIS-REGULATORY ELEMENTS ACROSS THE C-T BOUNDARY.....	82
4.1. Introduction.....	82

4.2. Results and Discussion	85
4.2.1. ATAC-Seq library preparation	85
4.2.2. Selection of Hox genes to investigate.....	87
4.2.3. Identifying and validating Hox <i>cis</i> -regulatory elements	89
4.2.3.1. Establishing methodology	90
4.2.3.2. Identification and validation of HOXC8 enhancer	92
4.2.3.3. Identification and validation of HOXC6 enhancer	96
4.2.3.4. Identification and validation of HOXC5 enhancer	100
4.2.3.5. Genome wide search for Hox cluster enhancers	101
4.2.4. Identifying downstream targets of HOXC6	103
4.3. Conclusion.....	105
CHAPTER 5: LONG NON-CODING RNAS	107
5.1. Introduction.....	107
5.1.1. Classification of lncRNAs.....	107
5.1.2. Identification of functional lncRNAs.....	109
5.1.3. Methods for investigating lncRNA functions	111
5.1.2. Identified functions of lncRNAs.....	113
5.2. Results and Discussion	117
5.2.1. Identification of lncRNAs from RNAseq dataset	117
5.2.2. Visualising lncRNAs in the chick embryo	118
5.2.3. A closer inspection of novel lncRNA ENSGALG00000053345.....	120
5.2.4. Interaction of novel lncRNA ENSGALG00000053345 with <i>HOXC6</i>	122
5.2. Conclusion.....	125
CHAPTER 6: IMPACT AND FUTURE PROSPECTS	127
REFERENCES	131
APPENDIX 1: PRIMERS AND GUIDES	158
APPENDIX 2: SEQUENCES.....	165
APPENDIX 3: RNA-SEQ.....	167

Figure Index

Figure 1.1: Somites and their derivatives	11
Figure 1.2: Gene expression patterns that are associated with paraxial mesoderm segmentation and maturation in the chick embryo	13
Figure 1.3: A schematic diagram showing the progressive differentiation of somites.....	16
Figure 1.4: Correspondence of Hox paralogues with vertebral formulae of mouse and chick embryos and primary/secondary body formation	18
Figure 1.5: Schematic representation of the somite levels bridging the cervical-thoracic transition in mouse, chick, goose, Xenopus, and zebrafish.....	20
Figure 1.6: Variations in Wing Position Correlate with Changes in Tbx5, Hoxb4, and Hoxb9 Expression Domains in the LPM (Lateral Plate Mesoderm).	20
Figure 1.7: Schematic representation of the four vertebrate Hox clusters.....	24
Figure 1.8: Conserved pattern of Hox gene organisation along the Drosophila cluster and their mammalian counterparts.....	28
Figure 1.9: Synpolydactyly.	31
Figure 1.10: Schematic representation of paralogous Hox axial skeleton phenotypes.	34
Figure 1.11: Comparison of Tbx5 and HoxC gene expression domains.....	36
Figure 1.12: Enhancer mode of action theory.....	39
Figure 2.1: pGEM-T Easy vector map.....	46
Figure 2.2: Schematic of heterotopic graft	52
Figure 2.3: Schematic of somite isolation procedure.....	53
Figure 2.4: Schematic of RNAseq/ATACseq experimental setup.	54
Figure 2.5: Mutagenesis PCR design.....	63
Figure 2.6: Guide RNA oligo design	65
Figure 2.7: U6.3 Mini Vector gRNA cloning site	65
Figure 3.1: A typical RNAseq experiment	68
Figure 3.2: Principal Component Analysis (PCA) Plots	75
Figure 3.3: Transcriptional Profiling of Somites across the C-T Boundary	77
Figure 3.4: RNA-Seq Data Quality Check and Visualisation.....	78
Figure 3.5: Hierarchical clustering of RNA-Seq data.....	80
Figure 4.1: ATAC-seq methodology	84
Figure 4.2: Quality control assessment of ATAC-Seq libraries.....	86
Figure 4.3: Hox gene expression pattern.	88
Figure 4.4: Ectopic HOXC6 expression following thoracic to cervical graft.	89
Figure 4.5: Initial attempts at candidate enhancer validation ex ovo	91
Figure 4.6: Identification of a regulatory element for HOXC8	93
Figure 4.7: Visualisation of a regulatory element for HOXC8.	94
Figure 4.8: Identification of a regulatory element for HOXC6.	97
Figure 4.9: Visualisation of a regulatory element for HOXC6	98
Figure 4.10: Identification of a regulatory element for HOXC5	100
Figure 4.11: Identification and validation of putative enhancers across Hox clusters A, B, and D.	102
Figure 4.12: A selection of candidate targets of HOXC6.....	104
Figure 5.1: Interactions of lncRNA HOTAIR to induce silencing to target loci	116
Figure 5.2: Whole mount in situ hybridisation of nine novel lncRNAs identified from the RNAseq dataset.....	119
Figure 5.3: Genomic location of novel lncRNA ENSGALG00000053345 in relation to the HOXC cluster	120
Figure 5.4: Comparison of ISH between HOXC6 and novel lncRNA	121
Figure 5.5: Visualisation of expression following CRISPR-ON of HOXC6.....	123
Figure 5.6: Visualisation of expression following CRISPR-ON of HOXC8.....	124

Table Index

Table 2.1: cDNA synthesis reaction setup	43
Table 2.2: cDNA amplification reaction setup	44
Table 2.3: cDNA amplification PCR conditions.....	44
Table 2.4: pGEM ligation reaction setup	45
Table 2.5: colony PCR reaction setup	47
Table 2.6: colony PCR conditions	47
Table 2.7: probe synthesis reaction setup.....	48
Table 2.8: Mastermix 1 reagents.....	54
Table 2.9: Mastermix 2 reagents.....	55
Table 2.10: Smart-seq2 PCR conditions	55
Table 2.11: Mastermix 3 reagents.....	56
Table 2.12: Smart-seq2 PCR conditions 2	56
Table 2.13: ATAC lysis/tagmentation buffer reagents	57
Table 2.14: ATACseq PCR conditions	57
Table 2.15: putative enhancer amplification reaction setup.....	59
Table 2.16: putative enhancer amplification PCR conditions	60
Table 2.17: BsmBI reaction setup.....	60
Table 2.18: BsmBI PCR conditions	60
Table 2.19: Enhancer mutagenesis reagents	63
Table 2.20: Enhancer mutagenesis PCR conditions.....	64
Table 2.21: sgRNA plasmid reaction setup	66
Table 2.22: sgRNA plasmid PCR conditions	66
Table 3.1: Samples for Differential 1 (D.1) of RNA-Seq	73
Table 3.2: Samples for Differential 2 (D.2) of RNA-Seq	74
Table 5.1: List of top 10 differential novel lncRNAs identified from RNAseq dataset.	117
Table 5.2: List of miRNA targets within novel lncRNA ENSGALG00000053345.	122

CHAPTER 1: INTRODUCTION

1.1. *The chick as a model organism*

The chick embryo (*Gallus gallus domesticus*) is a classic model for the study of development in a higher vertebrate species (Mok, 2015) and it is one of the first model organism to be used in developmental investigations (Stern, 2005). The chick embryo is extremely valuable, in part because almost the entirety of their development takes place within an egg that is incubated outside of the mother (Mok, 2015; Stern, 2005). As a result, early developmental stages can be accessed, visualized and manipulated by simply creating a small hole, or ‘window’, in the eggshell and embryos are easily accessible for *in vivo* imaging, manipulation of gene function, tissue dissection and cell sorting.

Billions of chickens are raised worldwide for meat and egg production already. Because of this, scientists can easily and economically acquire large numbers of fertilized eggs throughout the year. Furthermore, chickens share significant genetic conservation with humans. About 60% of chicken protein-coding genes have a single human orthologue (Hillier *et al.*, 2004). Therefore, at least some of the genetic mechanisms that have been found to regulate chick embryo development are also relevant to human biology.

1.2. *Development of the vertebral body plan*

1.2.1. Gastrulation

Gastrulation is a critical stage in the development of higher organisms, in which the body plan of the embryo is generated. In avian, reptilian, and mammalian embryonic development gastrulation starts with the formation of the primitive streak (Vasiev *et al.*, 2010). In the chick, the primitive streak forms during the initial stages of development at the embryonic midline. Streak formation is associated with large-scale cell flows that carry the epiblast cells overlying Koller's sickle into the central midline region of the embryo (Vasiev *et al.*, 2010). This movement is followed by ingression of cells, which undergo an epithelial-mesenchymal transition (EMT). This generates the three germ layers of the embryo: ectoderm, mesoderm, and endoderm (Solnica-Krezel & Sepich, 2012). Somites originate from the middle germ layer – the mesoderm.

1.2.2. Somitogenesis

Somites are paired aggregates of mesoderm cells that form sequentially on either side of the neural tube in a head-to-tail (anterior-posterior) progression in vertebrates. Their formation is coordinated with embryo growth at its posterior end (Casaca *et al.*, 2016; Nowicki & Burke, 2000) (Fig 1.1A). They are the earliest metameric structures that form in development (Kim *et al.*, 2011). As the primitive streak regresses and the neural folds begin to gather at the centre of the embryo, paraxial mesoderm, or presomitic mesoderm (PSM), separates into epithelial cells, which are rapidly patterned into distinct compartments that give rise to distinct cell lineages including muscle (Chal & Pourquie, 2017; Streit & Stern, 1999) (Fig 1.1B-D).

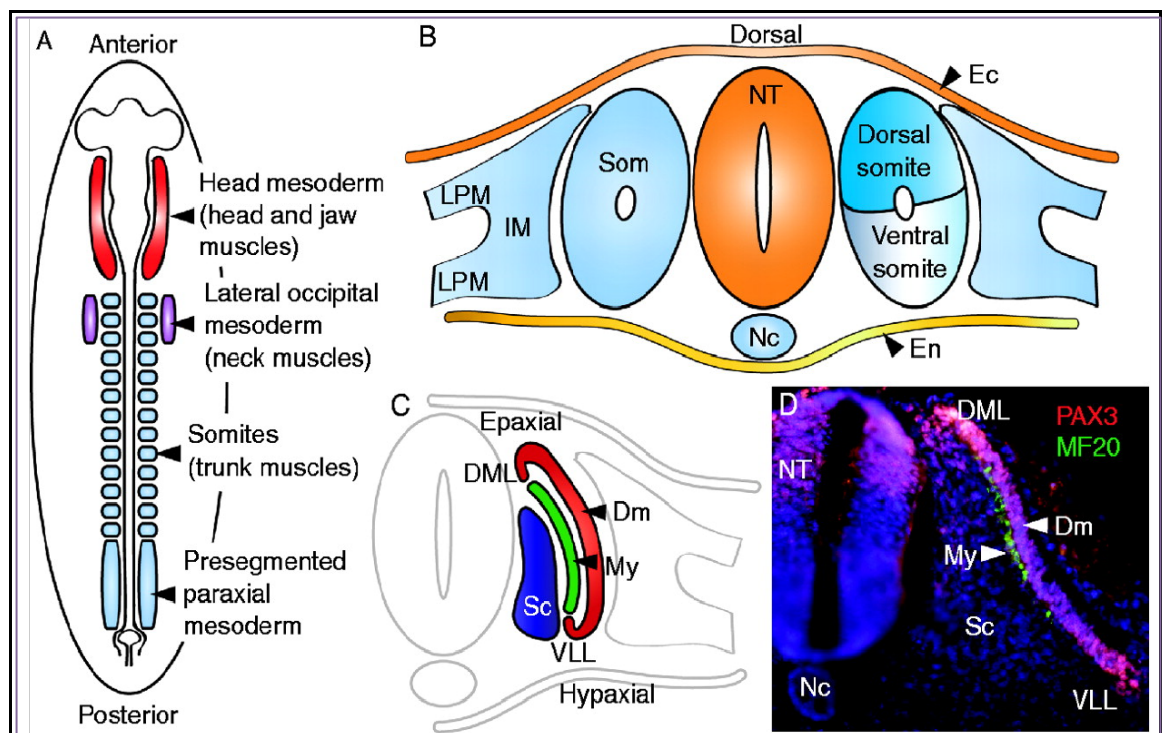


Figure 1.1: Somites and their derivatives. (A) Schematic of 13 somite pairs in an amniote embryo, corresponding to Carnegie stage 1 in humans (~24 day), E8.5 in mouse, and HH11 in chicken (40-45hr), showing the origin of different muscle groups. (B) Schematic of a transverse section of a chick embryo showing spatial arrangement of mesodermal derivatives in blue, ectoderm in orange and endoderm in yellow. The dorsal somite will produce the dermomyotome, and the ventral somite will produce the sclerotome. (C) Spatial organization of somite compartments as somites undergo differentiation: dermomyotome in red, myotome in green, and sclerotome in blue. (D) Immunostaining of transverse section of a differentiating somite in a stage HH18 chick embryo; PAX3 staining in red, MF20 (antibody to sarcomeric myosin, marker of differentiating muscle) in green, all nuclei are stained with DAPI in blue. LPM, lateral plate mesoderm; IM, intermediate mesoderm; Som, somite; NT, neural tube; Nc, notochord; Ec, ectoderm; En, endoderm; Dm, dermomyotome; My, myotome; Sc, sclerotome; DML, dorsomedial lip of dermomyotome; VLL, ventrolateral lip of dermomyotome. Figure taken from (Mok & Sweetman, 2011)

Somites are of fundamental importance as they are precursor populations of cells that will eventually differentiate into dermis, skeletal muscle, cartilage, tendons, and vertebrae, giving rise to the majority of the musculoskeletal system, including important structures such as limb and trunk muscles, the vertebral column, ribs, tendons and endothelial cells (Christ *et al.*, 2007). Paracrine signalling between somites and surrounding tissue can modulate development by determining the migratory paths of trunk neural crest cells (Kulesa & Gammill, 2010) and of the axons of spinal nerves (Rickmann *et al.*, 1985; Tannahill *et al.*, 1997).

1.2.3. Clock and wavefront model

Somites form at regular intervals, which are species-specific. The formation of the somites (somitogenesis) can be described by the “clock and wave” mechanism, first proposed by Cooke and Zeeman (1976). Here, an oscillating signal (the segmentation clock) provided by cascading genetic networks sets the pace of somite formation and interacts with a moving gradient of tissue competence in the PSM, which determines the position of the posterior border of the next somite (Baker *et al.*, 2006; Casaca *et al.*, 2016; Hubaud & Pourquie, 2014). A gradient of FGF8 is established along the body axis by mRNA decay (Dubrulle & Pourquié, 2004), interacting with the segmentation clock as it is displaced posteriorly to gate cells into somites (Baker *et al.*, 2006).

The segmentation clock consists of oscillating expression of Notch and Wnt signalling activities across the length of the PSM (Hubaud & Pourquie, 2014). Notch signalling, which is activated by Hairy and Lunatic Fringe (Lfng), promotes segmentation in the paraxial mesoderm, but this activity is silenced by Wnt and FGF signalling in the posterior PSM. When a new cycle of Notch activity reaches an area of the anterior PSM, where FGF and Wnt inhibitory signalling is below a threshold level, a molecular cascade is triggered, which acts to build a new segment border (Baker *et al.*, 2006; Casaca *et al.*, 2016; Hubaud & Pourquie, 2014) (Fig 1.2).

A gradient of Retinoic Acid (RA) exists going in the anterior-posterior direction, while the gradient of FGF and WNT go in the posterior-anterior direction (Aulehla & Pourquié, 2010). The region where these two gradients meet is called the ‘determination front’, or ‘wavefront’, and corresponds to the region where expression of oscillating genes is temporarily segregated and remains active to form a new pair of somites (Del Corral & Storey, 2004; Dubrulle & Pourquié, 2004).

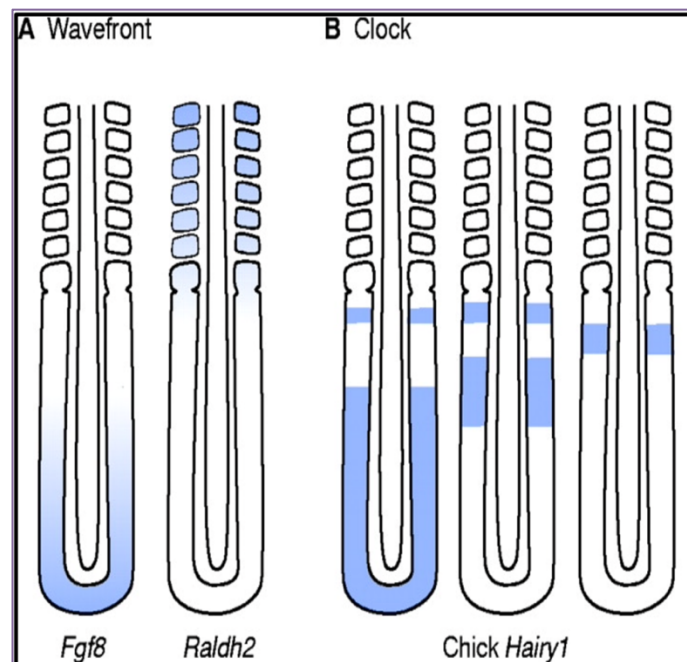


Figure 1.2: Gene expression patterns that are associated with paraxial mesoderm segmentation and maturation in the chick embryo. Schematic representations of gene expression patterns in the presomitic mesoderm (PSM) obtained by *in situ* hybridisation. **(A)** Genes that regulate the wavefront progression: the caudal-to-rostral Fgf8 (fibroblast growth factor 8) gradient (left) and the rostral-to-caudal Raldh2 (retinaldehyde dehydrogenase 2) gradient (right). **(B)** Expression of a cyclic gene, chick Hairy1, in three different phases that characterise the segmentation clock. Figure taken from (Dubrulle & Pourquié, 2004).

Recent studies questioned the role of this ‘clock and wavefront’ mechanism in somite formation and suggested that somites are self-organising structures. Palmeirim *et al.* (1997) demonstrated that the molecular segmentation of the anterior part of the presomitic mesoderm is an intrinsic property and that no signal coming from neighbouring tissues is required. In addition, Dias *et al.* (2014) showed when PSM is placed ectopically near extraembryonic regions, that cyclic expression of clock genes, as well as waves and gradients, are not necessary for somite formation. However, the ‘clock’ appears to be required for normal subdivision of the somites into anterior and posterior halves (Stern & Piatowska, 2015).

Somites are formed at specific and consistent intervals unique to each organism. In the chick embryo a new somite is formed every 90 minutes, and the somite pairs appear at the same time on either side of the notochord (Stern & Piatowska, 2015). Once the somite forms and matures, specification of different regions within each somite, in which the cells commit to certain cell fates occurs, resulting in the creation of a body plan for the organism.

1.2.4. Somite Compartmentalisation

The commitment to specific tissue types occurs relatively late in the development of the somite. When the somite is still immature and epithelial, it is plastic, meaning each of its cells can become any of the somite-derived structures. As the somite matures, it generates different compartments and cell lineages; sclerotome (cartilage), syndetome (tendons), myotome (skeletal muscle), dermatome (dermis), and endothelial cells (DeRuiter, 2010; Hirano *et al.*, 1995) (Fig 1.3).

During early maturation, the ventral portion of the somite undergoes an epithelial to mesenchymal transition (EMT) and epithelial somite formation. The sclerotome is formed by de-epithelialization of the ventral portion of the epithelial somite in response to notochord-derived signals (Fig 1.3). The sclerotome, along with mesenchymal cells from the somitocoel, will contribute to formation of most of the axial skeleton, ribs, cartilage, and connective tissues (Syndetome), and bones (vertebrae) (Scaal & Christ, 2004; Brent *et al.*, 2003; Baykal & Korkusuz, 2016).

The dorsal portion of the somite, now known as the dermomyotome, remains as an epithelial sheet (Maroto *et al.*, 2012). The dermomyotome further sub-divides into a medial portion, the epaxial domain, and a lateral portion, the hypaxial domain. The epaxial domain will give rise to the axial musculature and skeleton, while the hypaxial domain will contribute to the muscles of the limbs and body walls. The central part of the dorsal layer of the dermomyotome (also called the dermatome) will give rise to the muscles and dermis of the back (Ordahl & Le Douarin, 1992; Baykal & Korkusuz, 2016). The formation of the sclerotome and the dermomyotome define the dorsoventral axis of the somite.

There are several genetic signals that trigger the formation of the sclerotome and the dermomyotome. Sonic hedgehog (Shh) and noggin signals from the floor plate and the notochord promote de-epithelialisation and a mesenchymal fate, forming the sclerotome in the ventral part of the somite (Borycki *et al.*, 1998). Exposure to Shh induces Pax1 and Pax9 expression in the sclerotomal cells, which then triggers a genetic programme for further specification (Yusuf and Brand-Saberi, 2006). The sclerotome cells migrate to near the notochord and later produce the vertebrae and the ribs (Scaal and Christ, 2004). Shh signals from the notochord are also sufficient and necessary to activate MyoD gene expression in the myotomal region, beginning the muscular differentiation pathway (Borycki *et al.*, 1998). Canonical Wnt ligands, such as Wnt4, Wnt6 and Wnt8c signal from the overlying ectoderm and have been recognised as major dorsal signalling factors in the establishment of the epithelial

dermomyotome. They maintain the epithelial organisation of the dorsal part of the somite, inducing expression of Pax3 and Pax7, forming the dermomyotome (Münsterberg *et al.*, 1995, Yusuf and Brand-Saberi, 2006).

Cells then delaminate from the dermomyotome edges and migrate underneath to form the myotome, a third compartment, sandwiched between the remaining dermatome and sclerotome (Fig 1.3). The dorsal Wnt signals and the ventral Shh work synergistically to develop the myotome, and begin the muscular differentiation pathway (Münsterberg *et al.*, 1995; Borycki *et al.*, 1998). Cells from the medial and lateral parts of the dermomyotome begin to express the myogenic factors Myf5 and MyoD, giving rise to the underlying myotome, which goes on to produce the epaxial and hypaxial muscles. Canonical Wnt signalling is important in the myotome (Schmidt *et al.*, 2000) and can directly induce genes such as Paraxis/Tcf15, which maintains the epithelial organisation of the dorsal somite. Shh signals from the notochord with Wnt signalling regulate MyoD gene expression through Gli2 and Gli3 in the myotomal region (Borycki *et al.*, 2000). As the formation of the myotome requires both dorsal and ventral signals, it has been suggested that Shh and Wnt family members act through a concentration gradient that is established along the dorsoventral axis. This gradient is also maintained by the presence of competitive Wnt and Shh inhibitors, present in the prospective sclerotome and dermomyotome respectively. For example, the secreted frizzled-related protein 2 (Sfrp2) is a Wnt antagonist upregulated by Shh. It is expressed in the sclerotome, induced by Shh, and prevents the dorsalisng effect of Wnt signalling (Lee *et al.*, 2000). Only one candidate Shh antagonist has been discovered, the Growth-arrest specific gene1 (Gas1), which is inducible by Wnt proteins, (Lee *et al.*, 2001). Once the myotome has separated from the dermomyotome, Neurotrophin 3 (NTF3) expression maintains the dermatome, which gives rise to dermis tissue (Brill *et al.*, 1995).

Bone morphogenetic proteins, (BMPs) are another key family of proteins that work to pattern the somite. They play a variety of differing roles in the patterning of the somite. BMP4 is expressed in the dorsal neural tube where it indirectly induces the expression of Wnt11 in the dorsomedial lip of the dermomyotome (Marcelle *et al.*, 1997). This is the region that then forms the myotome and begins to express MyoD. However, other work suggests that BMP4 negatively regulates myogenesis, by preventing expression of MyoD and Myf5 (Reshef *et al.*, 1998, Pourquie *et al.*, 1995). Noggin antagonises BMP proteins by preventing them from binding their receptors (Zimmerman *et al.*, 1996) and it is present in the medial somite. This work suggests a dual role for BMP-4, depending on where its expression is present. When present in the dorsal neural tube, it indirectly induces expression of wnt-11 in the dermomyotome. However, when BMP4 is expressed in the lateral plate mesoderm, it induces the differentiation of

dermomyotome cell into Met expressing migratory muscle progenitor (MMP) cells, they continue to express Pax3, and myogenic differentiation is repressed.

The ventral sclerotome then gives rise to the syndetome, in response to myotomal signalling (Brent *et al.*, 2003) and the somite is then described as a fully differentiated somite. These signalling pathways act to produce and differentiate somites from PSM from stage HH7 to stage HH25, where somitogenesis is terminated (Tenin *et al.*, 2010). At limb levels, cells from the ventrolateral edges of the dermomyotome lose their epithelial characteristics, delaminate and migrate into the limb buds. These MMPs express the transcription factor ladybird (Lbx1) (Alvares *et al.*, 2003).

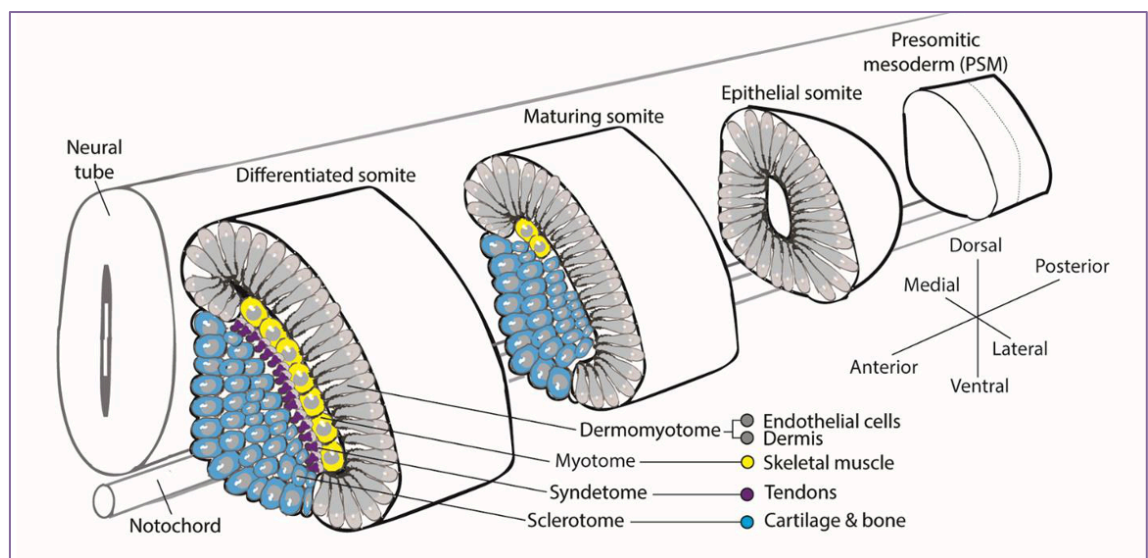


Figure 1.3: A schematic diagram showing the progressive differentiation of somites. The most posterior region is the presomitic mesoderm (PSM), which generates a new pair of epithelial somites. The somite then begins to mature, with ventral regions undergoing an EMT to give rise to two different compartments of the somite, the dermomyotome and the sclerotome. The most differentiated somite is the most anterior one, containing the sclerotome, syndetome, myotome and dermomyotome. These different regions give rise to different cell lineages (Adapted from Buckingham *et al.*, 2003).

The processes that follow somitogenesis include myogenesis (generation of muscle), tendon formation, and specification of the intermediate mesoderm (i.e., urogenital system) (Gilbert, 2000). Any disruption of the normal sequence of events in the formation or segmentation of somites can result in anomalies such as congenital vertebral defects (Chaturvedi *et al.*, 2018). Many spinal defects are also associated with kidney problems because the same lateral portions of the paraxial mesoderm cells that form somites also contribute stromal cells to the developing

kidney (Guillaume *et al.*, 2009). Somites are therefore critical in the planning of the vertebral body plan – essential for segmentation, bone and musculature development, as well as creating a template for the nervous system (DeRuiter, 2010).

1.3. *The Vertebral Axial Skeleton:*

The below section has been published in a review (Weldon & Münsterberg, 2022) but was expanded for more detail for this PhD thesis.

Somites are initially similar in morphology, but they produce anatomical features characteristic for their position along the body axis. Skeletal elements include vertebrae with complex morphology (Scaal, 2016), as well as scapulae and clavicles comprising the forelimb girdle, and pelvic bones comprising the hindlimb girdle. Based on these features, the axis is partitioned into six domains: occipital, cervical, thoracic, lumbar, sacral and caudal (Fig 1.4A, B). Transition from cervical to thoracic domains is associated with emergence of forelimbs, while the lumbar-sacral transition is associated with emergence of hindlimbs.

Within species the number of vertebrae in each domain is fixed, but between species the number of vertebrae in each domain varies, generating a species-specific axial formula (Iimura *et al.*, 2009). The occipital domain forms the bony structure at the base of the skull (Couly *et al.*, 1993; Huang *et al.*, 2000b); the cervical domain comprises neck vertebrae, linking vertebral column and skull; the thoracic domain is characterised by ribs and is followed by lumbar and sacral domains. The caudal domain, post-sacrally, represents the tail in most mammals and the rudimentary coccyx in humans. In birds, the caudal area includes a few small vertebrae and the pygostyle, formed by fused vertebrae (Schoenwolf, 1981; Rashid *et al.*, 2018; Rashid *et al.*, 2020). Distinctive molecular networks, including combinatorial Hox gene expression (see section 1.4.) produce these anatomical domains (Fig 1.4C).

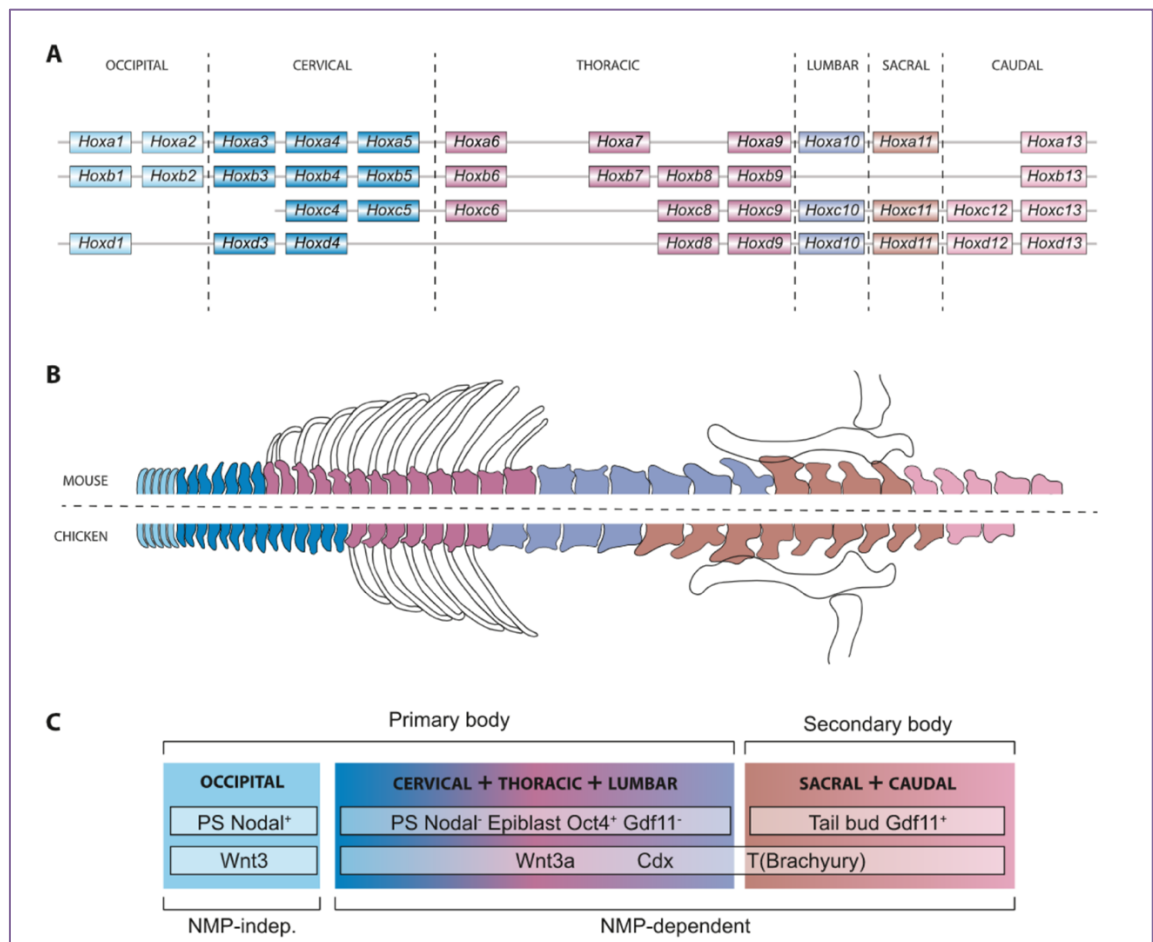


Figure 1.4: Correspondence of Hox paralogues with vertebral formulae of mouse and chick embryos and primary/secondary body formation. (A) Schematic of the genomic organisation of the four Hox clusters (a–d). Hox gene combinatorial expression controls vertebral identity along the anteroposterior axis. Parologue genes are colour-coded based on their contribution to axial domains (light blue = occipital; dark blue = cervical; dark pink = thoracic; purple = lumbar; brown = sacral; light pink = caudal). Boxes placed perpendicularly depict paralogous relationships within the Hox clusters. (B) Schematic representation of the vertebral formula of the mouse compared to the chicken. The somite-derived vertebrae are aligned and subdivided into colour-coded domains corresponding to axial domains, which correlate with specific Hox paralogue expression (A). (C) Schematic of primary and secondary body formation and the mechanisms for each. The colour-coded gradient corresponds to the vertebral axial formula in (B) and with specific Hox paralogue expression in (A). The occipital, cervical, thoracic and lumbar domains represent the primary body, whereas the sacral and caudal domains represent the secondary body. PS, primitive streak; NMP, neuromesodermal progenitor. Figure taken from (Weldon & Münsterberg, 2022).

In addition to conferring axial identity, Hox genes control the position of limb buds, forming from lateral plate mesoderm. In the forelimb, this is mediated via direct binding to a Tbx5 enhancer element. Co-electroporation of lacZ enhancer reporter plasmids with Hox expression constructs in chick showed that Hox binding sites are required to restrict Tbx5 expression to the forelimb-forming region. This occurs via both activating and repressive interactions. Similar observations were made using mice transgenic for lacZ enhancer reporters. For example, Hoxc9, which is

expressed in caudal lateral plate mesoderm, represses the *Tbx5* forelimb regulatory element (Nishimoto *et al.*, 2014). ChIP-seq experiments revealed overlapping binding sites for *Meis*, *Tbx*, and *Hox* in regulatory sequences of the *Fgf10* and *Lef1* genes. Thus, these transcription factors converge to co-regulate the Wnt and Fgf pathways, which are critical for limb initiation at the correct axial levels (Delgado *et al.*, 2021). Comparative analysis of multiple bird species: chicken, zebra finch and ostrich, with different numbers of cervical vertebrae, suggested that variation in forelimb position correlates with the timing of *Hox* gene activation during gastrulation (Moreau *et al.*, 2019; Young *et al.*, 2019).

Birds are one of the more diverse and widespread vertebrate orders, and they have long been a source of fascination to scientists (McPherson & Dodgson, 2004). The chicken (*Gallus gallus*) has 52 pairs of somites which will generate the different specialised regions of the vertebral column, occipital, cervical, thoracic, lumbosacral, and coccygeal (Gomez & Pourquié, 2009). At each different axial level, each somite is characterised by the expression of unique combinations of *Hox* paralogues (Fig 1.4).

1.3.1. Differences in Axial Formula Between Species.

The number of segments, or vertebrae, included at each anatomical domain in the embryo is referred to as its axial formula (Nowicki & Burke, 2000). A common form of evolutionary variation between vertebrate taxa is the changes in the axial formula of differing species (Burke *et al.*, 1995). While the different numbers of segments that contribute to various regions of the anterior-posterior axis is fixed within species, it can be vastly different between species. This change in axial formula is commonly referred to as ‘transposition’ (Burke *et al.*, 1995).

Embryonic segments are counted from the anterior end. The figure below shows the embryonic segments of the first three specialised regions of the vertebral skeleton, the occipital, cervical and thoracic domains, as they appear in the embryo of a mouse, chick, goose, *Xenopus* and zebrafish (Fig 1.5). The thoracic vertebrae begin formation at somite number 12/13 in the mouse; somite number 19/20 in the chick; 21/22 in the goose; 17/18 in the zebra finch; 22/23 in the ostrich (Fig 1.6).

Hox group 6 genes are involved in the formation of the thoracic area owing to their unique rib-promoting properties (Casaca *et al.*, 2016). During the formation of vertebrae, the most anterior embryonic segments that express *HOXC6* mark the end of the cervical (neck) vertebrae and the beginning of the thoracic (rib) vertebrae (Fig 1.5).

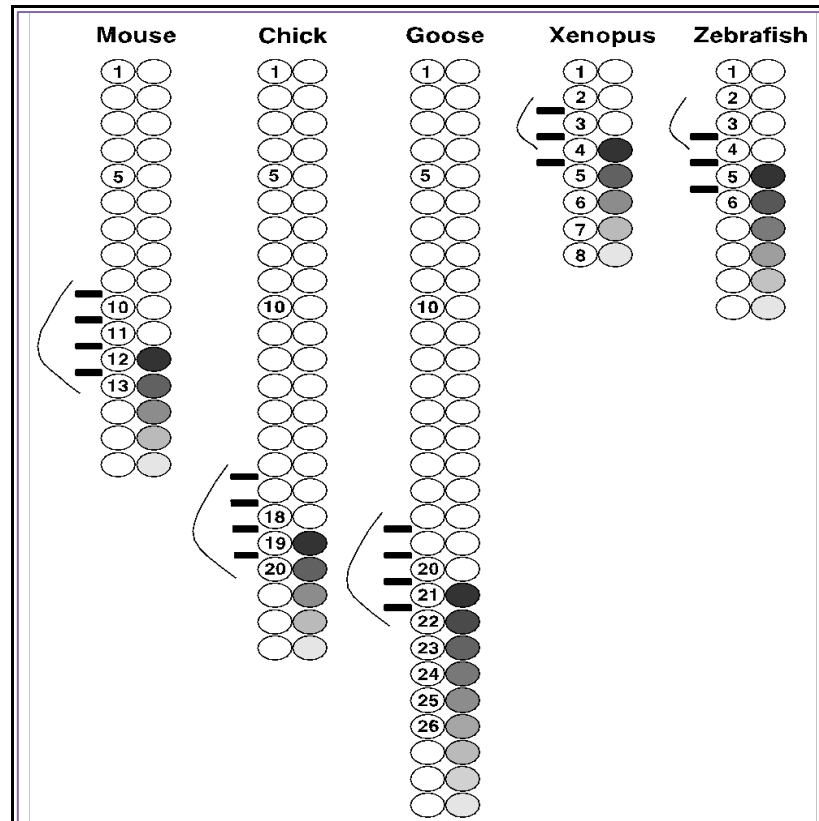


Figure 1.5: Schematic representation of the somite levels bridging the cervical-thoracic transition in mouse, chick, goose, *Xenopus*, and zebrafish. Black bars represent the spinal nerves of the brachial plexus, and the level of the limb or fin bud is indicated with a curved line. The shaded somite levels indicate the level of *HOXC6* expression, which marks the first thoracic segment, as determined by whole-mount *in situ* hybridisation, or immunohistochemistry. Figure taken from (Burke *et al.*, 1995).

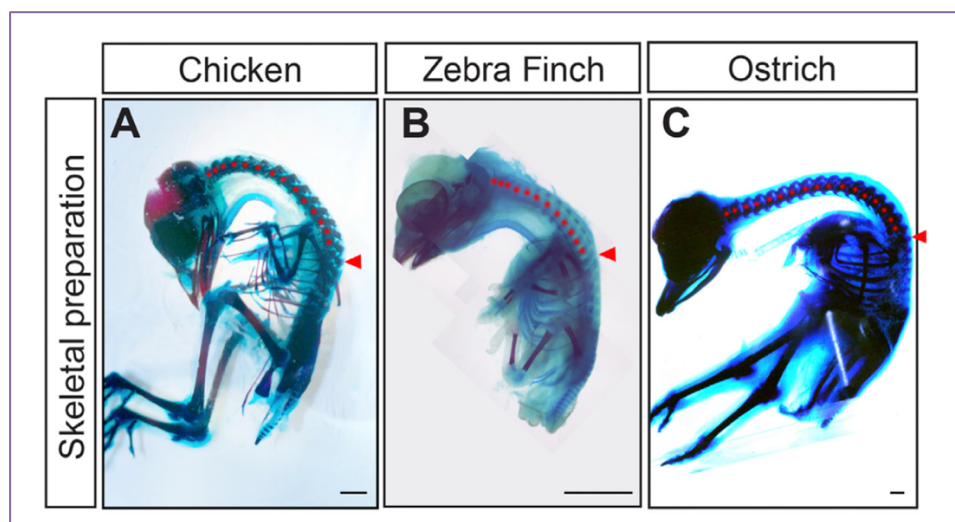


Figure 1.6: Variations in Wing Position Correlate with Changes in *Tbx5*, *Hoxb4*, and *Hoxb9* Expression Domains in the LPM (Lateral Plate Mesoderm). (A–C) Alcian blue-Alizarin red staining of chicken (A; E20), zebra finch (B; E13), and ostrich (C; E37). Red arrowheads point at the wing position level (15th, 13th, and 18th vertebrae in chicken, zebra finch, and ostrich embryos, respectively); red dots mark each cervical vertebra. Image adapted from (Moreau *et al.*, 2019).

Birds and mammals are the two groups of large-brained, endothermic (“warm-blooded”) animals on Earth. All mammals have seven cervical vertebrae, be it a mouse or a giraffe, but this number varies among the different species of birds. Birds have between 12-25 cervical vertebrae. The number of vertebrae in the neck varies from species to species in birds. However, the number of cervical vertebrae in mammals is remarkably constant, regardless of size, with virtually all mammals possessing seven. It is not clear why this number is fixed in mammals and not in birds, but there are some theories as to why this fixture persists.

Not only is the neck proportionally longer for birds than mammals, but it is also more flexible (Kardong, 2012). Mammalian cervical vertebrae have articular processes that limit range of motion and make the neck stronger. In birds, however, the equivalent processes are much smaller. Also, the shape of the vertebrae themselves allows for a greater range of motion. This allows birds to move their heads freely for feeding and flying and to reach any part of the body with the beak for grooming. Birds also have very rigid bodies with fused vertebral sections of the backbone (Reece *et al.*, 2014). This allows the bird to fly and the additional vertebral sections in the neck increases the much-needed range of movement of the bird. A longer neck also acts as a shock absorber when the bird lands after flight, as well as allows the bird’s centre of gravity to adjust when changing from a walking position, with the neck upright, to a horizontal position in flight (Kardong, 2012). These are some of the possible reasons as to why, evolutionarily birds have more and a varied number of cervical vertebrae than mammals.

It is also believed that at some point in time a lineage evolved where formation of cervical vertebrae became linked with the development of aspects of the cardiovascular system. This linkage did not develop in birds or reptiles, but it is this lineage that gave rise to mammals. Because of this link to the vascular system, any mutation that would affect vertebrate number in the cervical domain would also interfere with the circulation, particularly in the forelimbs and head, resulting in a reduction in survival of that organism, therefore making it near impossible for a species from this lineage to evolve any changes.

For example, thoracic outlet syndrome is a group of disorders that occur when blood vessels or nerves in the space between the collarbone (clavicle) and the first rib are compressed (NINDS, 2019). Thoracic outlet syndrome can be a result of an extra first rib (cervical rib), which can result in arterial thoracic outlet syndrome, caused by congenital (present at birth) bony abnormalities in the lower neck and upper chest. These can lead to poor circulation in the arms, hands and fingers. Therefore, a decrease in cervical number in a mammal such as the human, along with

an extra anterior thoracic segment (a rib), causes major disruptions to the cardiovascular system.

It has also been proposed that the lack of variation among mammals is caused by further developmental constraints. For example, changes in Hox gene expression that lead to changes in the number of cervical vertebrae are associated with neural problems and with an increased susceptibility to early childhood cancer and stillbirths (Galis, 1999). In humans and mice, Hox genes have been shown to be involved in the normal and abnormal proliferation of cell lines; several types of paediatric cancers are associated with abnormalities in Hox gene expression and congenital anomalies. In these embryonal cancers the incidence of a cervical rib (a rib on the seventh cervical vertebra, a homeotic transformation of a cervical vertebra towards a thoracic-type vertebra) appears to be increased. Birds and reptiles have a much lower susceptibility to cancer. This most likely reduces the deleterious pleiotropic effect of neonatal cancer when changes in cervical vertebral number occur. In mammals there is, thus, a coupling between the development of the axial skeleton and other functions (including cell proliferation). The coupling of functions is either a conserved trait that is also present in reptiles and birds, but without apparent deleterious effects, or the coupling is new to mammals due to a change in the functioning of Hox genes (Galis, 1999).

Something I would like to investigate with this project is if this coupling between the development of the axial skeleton and other functions is in fact also present within bird species or not. This information would be revealed through examination of the chick embryo, but further data may also be collected from additional bird species, such as the ostrich or zebra finch, in order to examine if this is possible among multiple bird species. This information could help provide some major insight into the evolution of separate animal lineages.

1.4. *Hox genes*

The successful organisation of the body plan requires that local information in the embryo be translated into a functional, global pattern. Hox genes are known to influence both the development and the evolution of this morphology (Nowicki & Burke, 2000). Metazoan development involves differentiation of an initial single cell into numerous different cell types, controlled by genetics and resulting in tissues such as bone, nerves and muscle. One would assume that along with the need for differential control during morphogenesis an embryo also needs a genetic control system to establish body asymmetry and planning. Transformation of

genetic signals in this instance would lead to positional information and identity for body structures (McGinnis & Krumlauf, 1992).

A genetic control system shown to regulate morphological planning during embryo development was discovered by Ed Lewis in 1978 after further investigation of a phenomenon initially described as homeosis (transformation of one body structure into that of another) by Bateson in 1894. This system of control is mediated by the "Homeobox genes", or Hox genes. Amniotes have nearly 40 members of the Hox gene family, which are expressed in staggered and overlapping domains along the embryonic axis (Chen *et al.*, 2013). They are organised into four gene clusters (Fig 1.4), which arose from two consecutive duplications of one ancestral cluster (Casaca *et al.*, 2016; Gomez & Pourquié, 2009). Due to these duplication events, each gene within a particular cluster is a close relative to genes from one or more of the remaining clusters which share a similar sequence and relative position within the cluster. These closely related genes were classified into 13 groups, known as paralog groups (Casaca *et al.*, 2016). The position of these genes along the cluster reflects the order in which each gene will be expressed in time and space along the anterior-posterior axis of the embryo (Gomez & Pourquié, 2009) (Fig 1.7).

Formation of the vertebrate axial skeleton requires the coordinated activity of multiple Hox genes, the spatial and temporal expression of which lead to differences along the anterior-posterior axis, thus defining their positional identity (Burke *et al.*, 1995). Hox proteins specify segmental identities in most animal embryos and their patterning roles in the vertebrate axial skeleton have been extensively characterised through phenotypic analysis in mice. However, the processes they regulate in somites and their derivatives – how they interact with somite patterning pathways to regulate different vertebral morphologies – are largely unknown (Chen *et al.*, 2013). This is what I aim to investigate in this thesis.

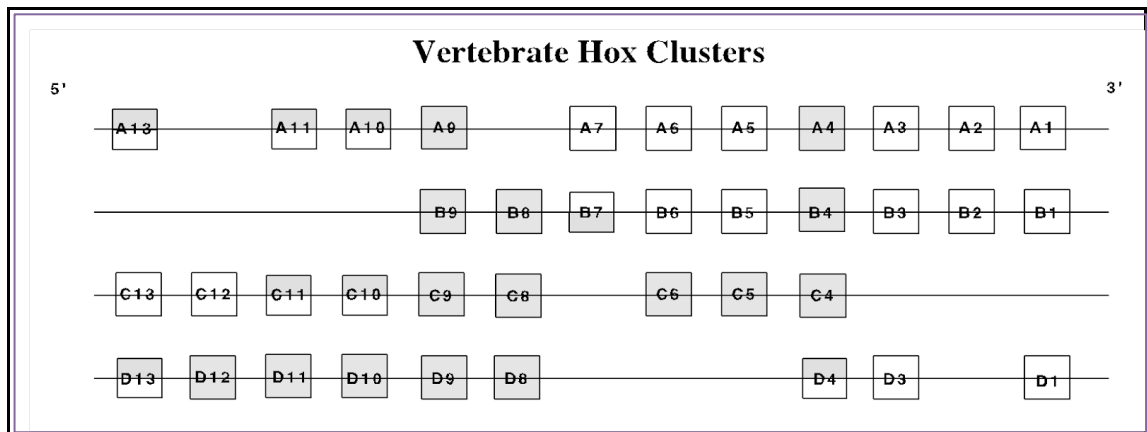


Figure 1.7: Schematic representation of the four vertebrate Hox clusters. The individual genes are represented by boxes. Filled boxes indicate genes examined in both the chick and the mouse. Half-filled boxes indicate genes examined in either the chick (top) or the mouse (bottom). Figure taken from (Burke *et al.*, 1995).

An important question is when axial identity becomes fixed and how this is achieved. Classic transplantation experiments in chick embryos can challenge differentiation potential (Kieny *et al.*, 1972). Grafting PSM between cervical and thoracic axial levels showed that the dorsal somite retains original Hox expression and morphological identity - specifically rib formation, vertebral and scapula shapes. However, laterally migrating cells are incorporated into ventral derivatives and adopt lateral plate specific Hox expression (Nowicki & Burke, 2000). Similarly, progenitor cells in the tailbud can adjust their Hox gene expression after heterochronic transplantation (McGrew *et al.*, 2008). This suggests that patterns of Hox expression are reversible and extrinsic signals from the environment can effect changes to intrinsic features, such as chromatin and epigenetic marks, to influence gene expression programmes.

Hox genes code for transcription factors involved in controlling the timing of ingression of paraxial (presomitic) mesoderm precursors into the primitive streak (Iimura *et al.*, 2009). In paraxial mesoderm tissue and in the adjacent neural tube, Hox expression is initiated by temporal activation of Hox genes from 3' to 5' within each cluster. This begins during gastrulation and is also detected in progenitor cell populations located caudally in the embryo (Wymeersch *et al.*, 2021). They possess a collinearity, where the precursors expressing more 5'-located Hox genes will ingress later than those expressing 3'-located Hox genes. The later ingressing cells will be found at a more posterior location. This results in a nested arrangement of the expression domain of these genes along the AP axis, with a specific combination of Hox genes expressed in each somite to specify the identity of each segment (Gomez & Pourquié, 2009; Wellik, 2007). Each somite pair differentiates into the cartilage, bone, muscle, and connective tissue of one vertebral segment and the type and unique morphology of each

segment depends on its location along the body axis (Chen *et al.*, 2013). It is likely that coexpression of multiple Hox genes at these specific locations controls specification of vertebral identities and initiates different morphologies.

Lineage tracing in chick showed that limited ingression and increased proliferation maintain and amplify the axial progenitor pool in the tailbud (Guillot *et al.*, 2021), and that these cells contribute to elongation of multiple axial tissues, for example neuromesodermal progenitors (NMPs) contribute cells to the neural tube and paraxial mesoderm/somites (Bénazéraf *et al.*, 2017). Transcriptomics, including single cell sequencing, identified their distinct molecular signatures (Guillot *et al.*, 2021; Wymeersch *et al.*, 2019). Their bipotency and eventual differentiation into mesodermal or neural progenies is determined by Brachyury-T and Sox2 transcription factors (Wymeersch *et al.*, 2021; Koch *et al.*, 2017). Furthermore, analysis of T-/- chimeric embryos suggests that Brachyury-T limits the allocation of primitive streak cells to the NMP pool (Guibentif *et al.*, 2021). Misexpression approaches in chick embryos provided evidence that the Sox2:Brachyury-T ratio affects the cells' motile behaviour: cells stay in place to generate neural tissue (high Sox2) or exit the progenitor zone to produce mesoderm (high Brachyury-T) (Romanos *et al.*, 2021). Transcriptomics identified a network involving RA and Wnt signalling regulating neural versus mesodermal allocation (Gouti *et al.*, 2017). The most highly upregulated genes in NMP- and lateral/paraxial mesoderm (LPMP)-containing regions were Hox genes, with peak expression during primitive streak to tail bud transition (Wymeersch *et al.*, 2019). The NMP transcriptome changes over time (Wymeersch *et al.*, 2019) and acquisition of axial identity remains responsive to extrinsic signals (McGrew *et al.*, 2008) until it becomes fixed in the PSM (Nowicki & Burke, 2000). Interestingly, anterior limits of Hox gene expression differ in paraxial mesoderm and neural tissues, despite their common origin from posterior progenitor pools (Wymeersch *et al.*, 2021) and the coupling of axial and paraxial tissue elongation (Xiong *et al.*, 2020).

As these progenitors contribute to more caudal mesodermal or neural progenies, collinear Hox gene activation is translated into a spatial pattern of expression along the axis (Mallo *et al.*, 2010). The progressive opening of Hox clusters is associated with changes in chromatin structure (Soshnikova & DuBoule, 2009). In chick this has been visualised using chromatin accessibility and Hox expression within PSM and somites. Differentially accessible chromatin detected across all four Hox clusters reflects the organisation of genes within clusters and their collinear expression. Footprints for transcription factors involved in Hox gene regulation and patterning, such as CDX1/2 and RA receptors, were found in intergenic regions (Mok *et al.*, 2021).

Correct Hox gene expression is not only important along the anterior posterior axis but also within the patterning of organ systems, where multiple Hox gene expression appears to set up a Hox code specifying and maintaining organ identity, such as the limb buds. Understanding all aspects of Hox gene regulation will help uncover the importance of their conserved clustered arrangement and the genetic interactions behind their expression patterns and establishment of morphological phenotypes, which vary between species.

1.4.1. Discovery of Hox Genes

Bateson initially described homeosis in 1894 as a phenomenon where, "a change occurs and one body part takes on the characteristics of another," which was at the time unexplainable genetically (Reviewed by Veraksa *et al.*, 2000). Lewis then went on to discover a genetic control system functioning along the A-P body axis offering, an explanation for Bateson's observations.

In 1978 Ed Lewis made a large scale and schematic study mapping mutations resulting in homeotic transformations in *Drosophila* (Mann, 1997). One such example of these homeotic mutations is the dominant mutation of the Antennapedia (Antp) gene. The Antp gene is thought to specify identity in the second thoracic segment with a dominant mutation causing ectopic expression in the head as well as the thorax resulting in legs developing from the head socket rather than antennae (Scott *et al.*, 1984). Another well-known homeotic mutation is that of Ultrabithorax (Ubx) again classified as a transformation where halteres, the small club-like balancing organs of flies, transform into an extra pair of wings.

Primarily identified in *Drosophila* but later observed in all bilateral animals, the control system for this was then shown to be through a conserved and clustered arrangement of genes "Hox genes" and the proteins encoded by them. Hox genes are expressed in a conserved pattern in specific body segments and co-linearity is shown between the order of genes along the chromosome and the order in which the Hox genes are expressed, giving some explanation to the co-linearity and transformation of body structures by mutation to genes within this complex.

The *Drosophila* Hox cluster (HOM-C) contains a total of 10 genes, is split into two parts and located on the same chromosome; correct expression of the Hox genes was demonstrated to be paramount for the correct development of *Drosophila*. Mutations within the cluster could cause segments of the body to be converted into that of another, a phenomenon referred to as homeosis, leading to the genes being labelled "Master Regulators", due to the control they exert over the resulting phenotype (Lewis, 1978).

The sequencing of the first three Hox genes in *Drosophila* showed they share a Homeobox found in regulatory genes, proteins made from these genes contain a 60 amino acid stretch, the Homeodomain. The Homeodomain proteins function as transcription factors which, activate or repress multiple downstream target genes (Zhao *et al.*, 2002; McGinnis & Krumlauf, 1992; Scott, 1993). The structure of the homeodomain can be related to that of the helix-turn-helix motif of prokaryotic DNA binding proteins e.g. *Lac repressor*.

Subsequent research carried out using several different species, ranging from *Drosophila* to mouse, and investigations into their development has shown genetic control systems working in the same way by the same principals conserved between phyla (McGinnis & Krumlauf, 1992).

1.4.2. Vertebrate Hox genes

Vertebrate Hox genes provide spatial and temporal cues for development during morphogenesis with the same co-linearity between the position of genes along the chromosome and their order of expression, as first identified in *Drosophila*. Regulation of expression is tightly linked to 3' to 5' order (Fig 1.7). The most 3' genes expressed first proximally/anteriorly and the 5' genes later and more distally/posteriorly along the A-P body axis (McGinnis & Krumlauf, 1992; Gerard *et al.*, 1993). This co-linear expression has also been observed in limb pattern formation and along other body axis (Gerard *et al.*, 1993; Scott, 1993; Mann, 1997).

Hox genes not only control developmental positioning but are also involved in the control of cell differentiation as shown by studies into congenital, somatic and metabolic alterations involving a variety of Hox genes in a range of different organs and tissues, such as the limb buds or the bladder (Cantile *et al.*, 2003). Specific groups of Hox genes or a Hox code is involved in correct developmental identity and positioning as well as maintenance of specific organ phenotypes. Correct Hox code is paramount to correct and complete organ function before and after birth. This is shown in recent studies where normal and carcinomal tissues have been analysed for their specific Hox code, which revealed differing expression codes between the two tissue types (Cantile *et al.*, 2003).

In vertebrate genomes class 1 homeobox genes, homologous to HOM-C in *Drosophila* (Fig: 1.8), comprise a family of 39 clustered genes with the estimated total number of homeobox containing proteins now being 200 if all the dispersed/non-clustered Hox genes are included (Zhao *et al.*, 2002). The complexes of about 100kb are defined as Hox loci, conveniently named HoxA located on human chromosome 7 (7p14-21), HoxB located on chromosome 17 (17q21-22), HoxC located on chromosome 12 (12q12-13) and HoxD located on chromosome 2 (2q.31-37) (Apiou *et al.*, 1996; Chiba 1998). Nomenclature refers to these complexes as Hoxa in mice and

HOXA in humans, a system approved by both nomenclature committees in 1993 (Boncinelli, 1997; Scott, 1993).

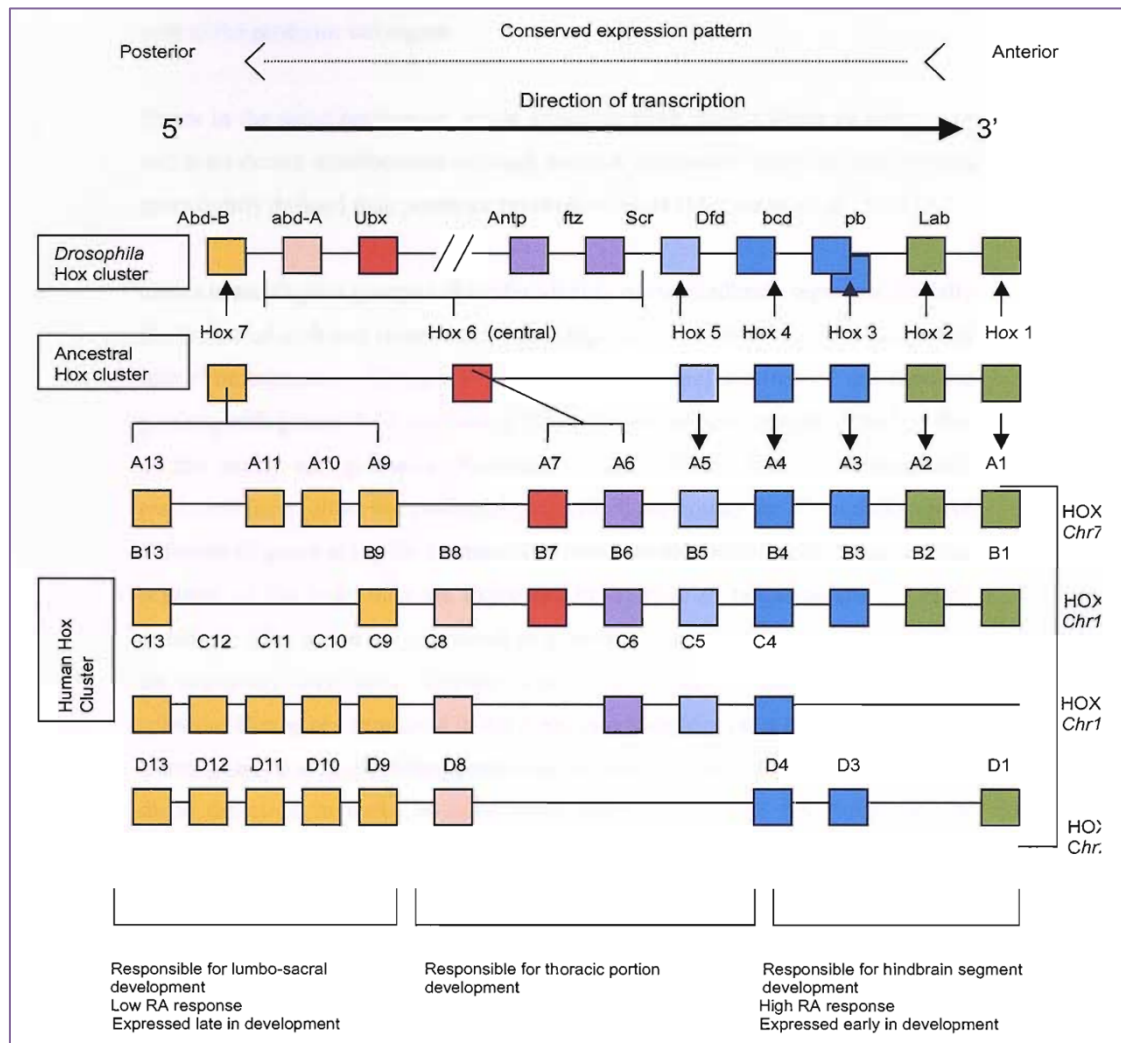


Figure 1.8: Conserved pattern of Hox gene organisation along the Drosophila cluster and their mammalian counterparts. Adapted figure from (Veraksa *et al.*, 2000).

Figure 1.8 shows the 39 human Hox genes in four clusters, each cluster containing between 9 and 11 genes apparently arising from two duplications of a primordial cluster. These genes collectively determine the positional identity of the various body segments throughout embryonic development, from the anterior branchial arch, at hindbrain level, to the posterior tail region.

Genes in the same paralogous group generally have similar limits of expression and roles during development although anterior expression limits of Hox genes is more tightly defined than posterior expression limits (McGinnis & Krumlauf, 1992).

Genes in paralogous groups 1-4 confer identity of the hindbrain segments, initially the branchial arch and rhombencephalon regions of the embryo, organizing them into rhombomeres. Groups 5-8 take developmental control of the thoracic portion, with groups 9-13 overseeing the lumbo-sacral development including that of the genitalia (McGinnis & Krumlauf, 1992). These developmental responsibilities follow the co-linearity described previously for *Drosophila* where the order of genes along the chromosome relates to the order of expression and the segment of the body they are expressed in, anterior to posterior and 3' to 5' direction. Hox genes are expressed in a variety of developing tissues, including the secondary axis limbs. Tetrapod limbs arose relatively late in evolution, the subset of Hox genes expressed in the limbs may therefore exert the same functions in the primary axis providing positional information and identity in a proximal distal direction in limbs or may have different roles and functions entirely (McGinnis & Krumlauf, 1992; Morgan, 1996; Burke *et al.*, 2001).

An interesting phenomenon relating to transformations due to a change in Hox gene expression pattern is referred to as "posterior prevalence". When a posterior Hox protein is expressed ectopically in an anterior domain the domain is transformed to have posterior characteristics, this effect however is not observed when anterior Hox proteins are expressed posteriorly as no transformations are observed. This phenomenon is observed in rostrocaudal patterning of the *Drosophila* embryo and the vertebrate hindbrain. This effect and the distinct functional roles for Hox proteins may be as a result of co-operative DNA binding with other protein cofactors, which in turn enhance the specificity of the proteins giving each one a distinct function (LaRonde-LeBlanc, 2003).

It is not fully understood how the Hox gene system has remained so highly conserved between phyla when there are such huge differences in their phenotypic appearances. Could these differences be due to different downstream targets or the mechanism of Hox gene regulation? Theories put forward in respect to Hox genes and morphological changes in evolution include the following four: 1) There are changes in Hox gene transcription patterns within a said segment 2) Hox protein responsive elements of downstream genes change 3) Hox expression patterns in various segments may vary and finally 4) The Hox gene in an organism can change (Krumlauf, 1992).

1.4.3. Interaction of Hox with lineage specifying genes

Ultimately Hox expression instructs anatomically distinguishable vertebrae subtypes of the vertebral column (Fig 1.4A, B) (Chal & Pourquié, 2017; Scaal, 2016; Scaal, 2020). Our understanding of how Hox genes interact with somite patterning pathways to regulate vertebral morphologies is incomplete. For example, cervico-thoracic vertebrae are specified by Hox paralogous groups 5-9 (Mallo *et al.*, 2010). *Hoxa5* is expressed in posterior cervical somites, where *Fgf-8* and *Shh* signalling restricts expression to the lateral sclerotome. Cartilage defects in lateral vertebral elements were observed after *Hoxa5* knockdown or misexpression and this is mediated by negative regulation of the chondrogenic transcription factor, *Sox-9* (Chen *et al.*, 2013). Similarly, *Hoxc6* expression demarcates the anterior boundary of the thoracic region in many vertebrates including mouse and chick (Gaunt, 1994; Burke *et al.*, 1995), alligator (Mansfield & Abzhanov, 2010) and snakes (Woltering *et al.*, 2009). Overexpression of *Hox6* paralogues in PSM produces ectopic ribs in more anterior cervical and more posterior lumbar regions, indicating that *Hox6* can determine thoracic identity throughout the vertebral column. Mouse knockout studies showed that *Hox10* and *Hox11* paralogues provide identity to lumbo-sacral regions. In absence of *Hox11* function, sacral vertebrae assume lumbar identity, whereas in absence of *Hox10* function, ribs project from all posterior vertebrae and no lumbar vertebrae form (Wellik & Capecchi, 2003). Conversely ectopic expression of *Hoxa10* suppresses rib formation (Carapuço *et al.*, 2005; Vinagre *et al.*, 2010). Interestingly, both *Hox6* and *Hox10* interact with an enhancer element that regulates expression of *MYF5* and *MRF4/Myf6* in the hypaxial myotome, which in turn activates FGF and PDGF signalling to promote rib formation (Vinagre *et al.*, 2010). Some species, including snakes, have a polymorphism in this enhancer element modulating its response to rib-suppressing and rib-promoting Hox proteins, overall resulting in an expanded rib cage (Guerreiro *et al.*, 2013).

My thesis aimed to further characterise downstream gene regulatory networks underlying axial transitions and skeletal morphology using differential transcriptomics and chromatin accessibility in chick somites. Focussing on cervical and thoracic levels, the objective was to identify molecular signatures that correlate with differential expression of Hox paralogues, which define this region. This has uncovered additional candidate target genes and gene networks that I propose mediate axial specific somite patterning and differentiation downstream of Hox.

1.4.4. Genetic Evidence for Hox gene Function in Visceral Organs and Limbs

Vertebrate Hox genes not only confer A-P positional information and regulate axial development. Overlaps in the expression of multiple Hox genes also sets up an apparent Hox code to specificity and maintain organ specific identity.

Hox genes expressed in the developing limb buds and external genitalia such as the posterior HOXD and HOXA genes are, important for the correct development of distal limb and genital structures and mutations can lead to so obvious phenotypic alterations (Veraksa *et al.*, 2000). A good example of such mutations is hand-foot-genital (HFG) (Stern *et al.*, 1970) syndrome, caused by heterozygosity for a non-sense mutation in the homeodomain of HOXA13 (Veraksa *et al.*, 2000).

Limb anomalies in HNF patients include short first metacarpals of normal thickness, small distal phalanges of the thumbs, short middle phalanges of the fifth fingers, and fusion or delayed ossification of wrist bones. Foot abnormalities include a shorter big toe due to a short first metatarsal and a small, pointed distal phalanx. Uterine anomalies in females with HFG include partial or complete division of the uterus, representing defects of Mullerian duct fusion. Urinary tract malformations in HFG females include a displaced urethral opening and malposition of ureteral orifices in the bladder wall; affected males may have hypospadias a misplacement of the urethral opening.

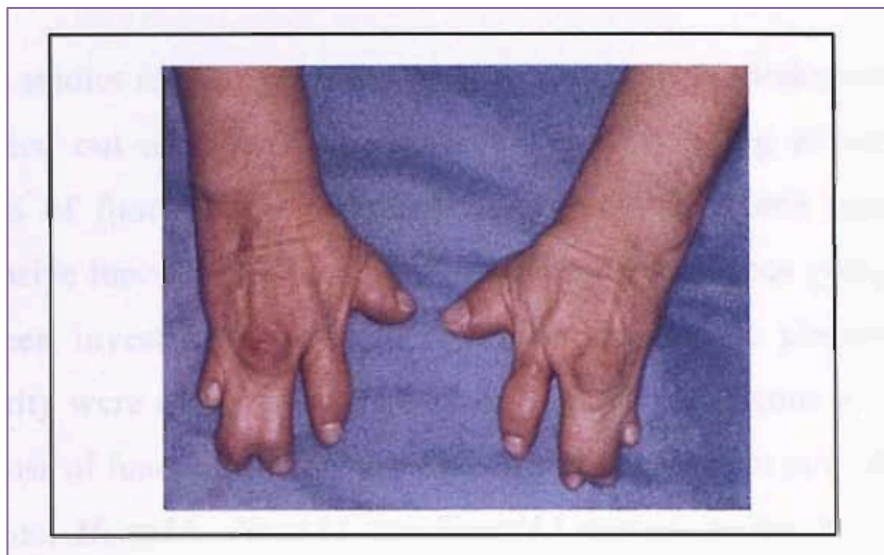


Figure 1.9: Synpolydactyly. This is due to expansion in a 15 residue polyalanine. Resulting in malformed hands and feet Figure taken from (Gilbert, 2000).

Heterozygous and homozygous synpolydactyly phenotypes (SPD) (Anderson, 1898) co segregate with an expansion in a normally IS-residue polyalanine stretch in exon one of HOXD 13. The mutation causes a dominantly inherited congenital condition and increasing severity of the phenotype correlates with expansion time (Veraksa *et al.*, 2000). The mechanism here is malformation most probably due to gain of function in the hands and feet with heterozygous patients having shorter digits and central syndactyly. Homozygous patients have more severe phenotypes showing transformation of metacarpal and metatarsal bones to short carpal and tarsal like bones (Innis, 1997; Zakany *et al.*, 1999). The temporal and spatial collinearity of the Hox genes is highlighted here with the mutation in HoxD13 resulting in malformation in the distal structures i.e., the digits.

In addition to malformations Hox genes are increasingly seen to be involved in other types of disease. HOXA9 is reportedly involved in myeloid leukaemia in humans. This form of leukaemia is associated with a translocation between chromosomes 7 and 11, between the regions where HOXA and the nucleoporin gene NUP98 are located (Nakamura *et al.*, 1996). Chromosome rearrangement in this region has been shown to result in genomic fusion of HOXA9 and NUP98, thought to promote leukaemogenesis through inhibition of the terminal differentiated and/or aberrant nucleocytoplasmic in the effected blood cells (Boncinelli, 1997).

Loss of function studies for Hox genes either individually or as paralogous groups have been carried out in order to understand the functioning of the genes. Frequently, loss of function of individual Hox genes has been masked by apparently extensive functional redundancy between the paralogous groups. This has recently been investigated for the Hox11 paralog where phenotypes of increasing severity were observed when 1, 2 or 3 of the paralogous genes were knocked out. Loss of function studies during kidney development provided some interesting points, Hoxa11, Hoxc11 and Hoxd11 appear to be functionally redundant in early kidney development, but both seem necessary for initiation of metanephric development. Single and double loss-of -function phenotypes were less severe than triple indicating some amount of functional redundancy within the paralog (Wellik *et al.*, 2002).

Current interest is also strongly focused on analysis of the whole network of Hox genes and their expression patterns or the specific Hox code in different organs and tissues and drawing comparisons of these apparent codes between the normal and carcinomal samples. This has uncovered interesting differences and in bladder cancer such as kidney problems related to the Hox11 paralogous group, bladder cancer analysis has uncovered alterations in the Hox11 paralogous group expression patterns. HoxA11 is silent in bladder cancer samples but active in

most normal, HoxC11 is silent in normal and active in bladder cancer samples and HoxD11 always silent in normal and expressed in bladder cancers. Hyperexpression is observed with 4 out of the 6 HoxC genes expressed in bladder cancers when silent in normal samples. This would point to the HoxC genes and the Hox11 paralog being specifically involved in bladder cancers to some extent (Cantile *et al.*, 2003). This and other studies show an apparent difference in Hox codes between normal and cancer tissues indicating a connection between Hox code alterations and the onset of cancers. A further understanding of Hox gene expression in these circumstances could lead to a better understanding of the pathways leading to the onset of cancers and possible development of cancer treatments in the long term.

The likely involvement of Hox genes in a range of diseases needs further investigation, we need a better understanding of the roles, and regulation of the Hox genes during and after development, which should help fully understand the aetiology of such diseases. Incorrect expression patterns, regulation of Hox genes and mutations to specific genes or co-factors are all issues, which need to be further addressed to fully understand why these changes occur.

1.4.5. Homeotic transformations of the axial skeleton.

Amniote Hox mutations can produce homeotic transformations, in which segments develop with morphologies inconsistent with their axial levels (Chen *et al.*, 2013) (Fig 1.10). Hox expression begins during gastrulation and AP regional identities are established in the PSM. This early expression is associated with determination of segmental identities: misexpression of Hox genes in PSM (but not after segmentation) leads to homeotic transformation, and somites transplanted after segmentation retain the axial identity and Hox expression of their original location (Burke & Nowicki, 2001; Casaca *et al.*, 2016; Chen *et al.*, 2013; Kant & Goldstein, 1999).

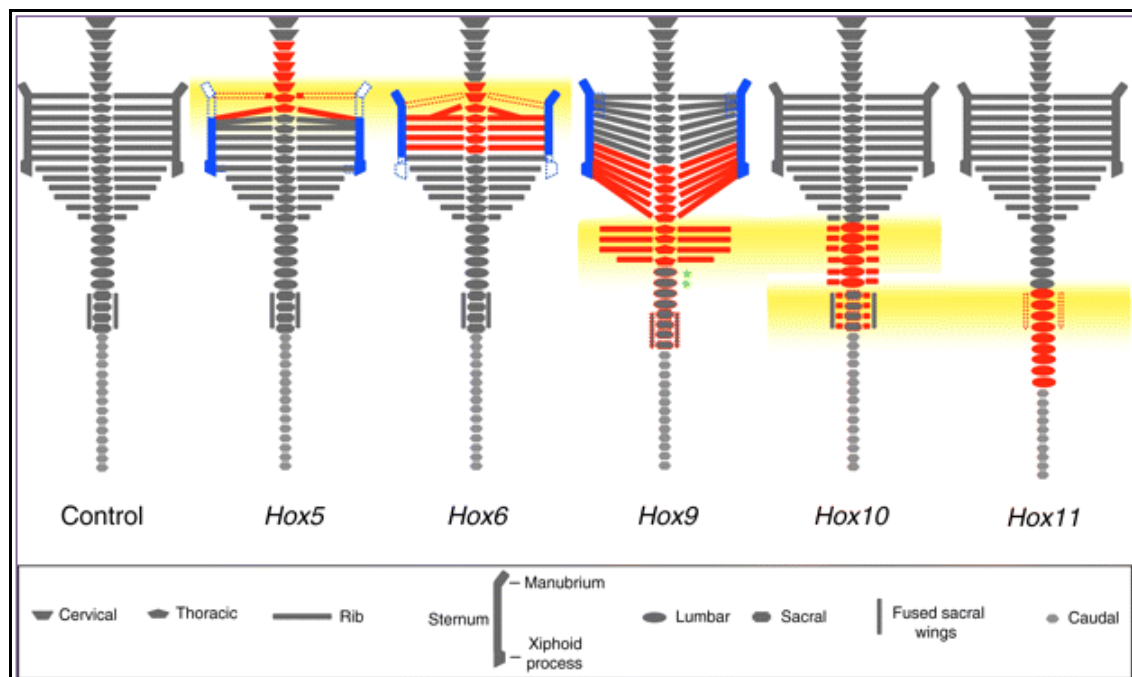


Figure 1.10: Schematic representation of paralogous Hox axial skeleton phenotypes. Somite-derived, primaxial skeletal elements that exhibit anterior homeotic transformations in paralogous *mice* mutants are in red. Lateral plate-derived structures affected in paralogous mutants are in blue. Note that Hox5, Hox6 and Hox9 mutant phenotypes are offset in their AP extent for defects in the somite-derived (red) and lateral plate-derived (blue) skeleton. The yellow shading highlights somite-derived AP regions that are affected in adjacent paralogous mutant groups. In each case, the overlapping regions of phenotype display distinct vertebral morphologies for each paralogous mutant. The green asterisks in the Hox9 mutant reflect the posterior shift of the axial skeleton. Mice normally possess 28-30 caudal vertebrae. Only 15 are represented here for simplicity. Despite changes in the number of caudal vertebrae in Hox9 and Hox11 paralogous mutants, the same average total numbers of vertebrae are present in all of the paralogous mutants depicted in this diagram and are the same as in control animals. Figure taken from (McIntyre *et al.*, 2007).

Hox gene expression domains often extend from their anterior expression in somites all the way to the end of the embryo, but their action is restricted to their anterior-most expression domain. Therefore, the posterior-most hox genes expressed control the identity of a segment. This mechanism is known as posterior dominance/prevalence (Gomez & Pourquié, 2009).

In contrast to the morphogenetic plasticity of the occipital somites, as described above, their characteristic inability to support survival of dorsal root ganglia was not significantly modified by posterior transplantation (Kant & Goldstein, 1999). Dorsal root ganglia, which are clusters of neurons in a spinal nerve, initially developed and then degenerated with the same morphological stages as normally observed, so this was a fixed phenotype, despite its ability to form caudal vertebrae after transplantation. In striking contrast to the plasticity of morphology, it was found that all four members of the fourth paralogous group of Hox genes that are expressed endogenously at the level of the graft are not upregulated in the caudal-transposed cranial mesoderm. It therefore appears that genes other than those of the Hox family normally

expressed at this axial level control the position-specific morphogenesis of ectopic vertebrae formed from cranial somites (Kant & Goldstein, 1999). This is further evidence of the presence of undiscovered genes that may confer identity to vertebral segments, which may be uncovered in this thesis.

1.4.6. Hoxc6, Hoxc5, and Hoxc8.

In vertebrates, Hox transcription factors are important for proper development of the axial formula described above. Specific Hox paralogues are related to segment type not segment number. For example, in chickens, cervical (neck) segments extend to somite 19, whereas in mammals they extend to somite 12. Yet regardless of the segment number, the change from cervical-type to thoracic-type is marked by the same transition of Hox gene expression (Fig 1.4A, B).

Hoxc6, a member of the sixth paralogue group, has an anterior boundary of expression at the level of the first thoracic vertebrae in both the chicken and the mouse (Fig 1.11). In the chicken, Hoxc6 is first expressed at somite level 19-20, two somites behind Hoxc5 expression, aligned with the middle of the forelimb bud. In the mouse Hoxc6 expression begins at somite level 12-13 and is also two somites behind Hoxc5 (Burke *et al.*, 1995) expression and in alignment with the middle of the forelimb bud (Burke *et al.*, 1995). Hoxc8 is another example of a Hox gene expressed throughout the thoracic domain (Fig 1.11). It is these three Hox genes that I will use as markers for cervical and thoracic level somites throughout my experiments. In order to confirm that a successful graft of thoracic segments into the cervical area has been carried out, Hoxc6 and Hoxc8 probes can be used to visualize the anteriorly grafted somites on the operated side of the embryo. These can then be compared to the wild type Hoxc6 expression on the non-operated side, and also confirm that no unintentional changes have been made to that non-operated side.

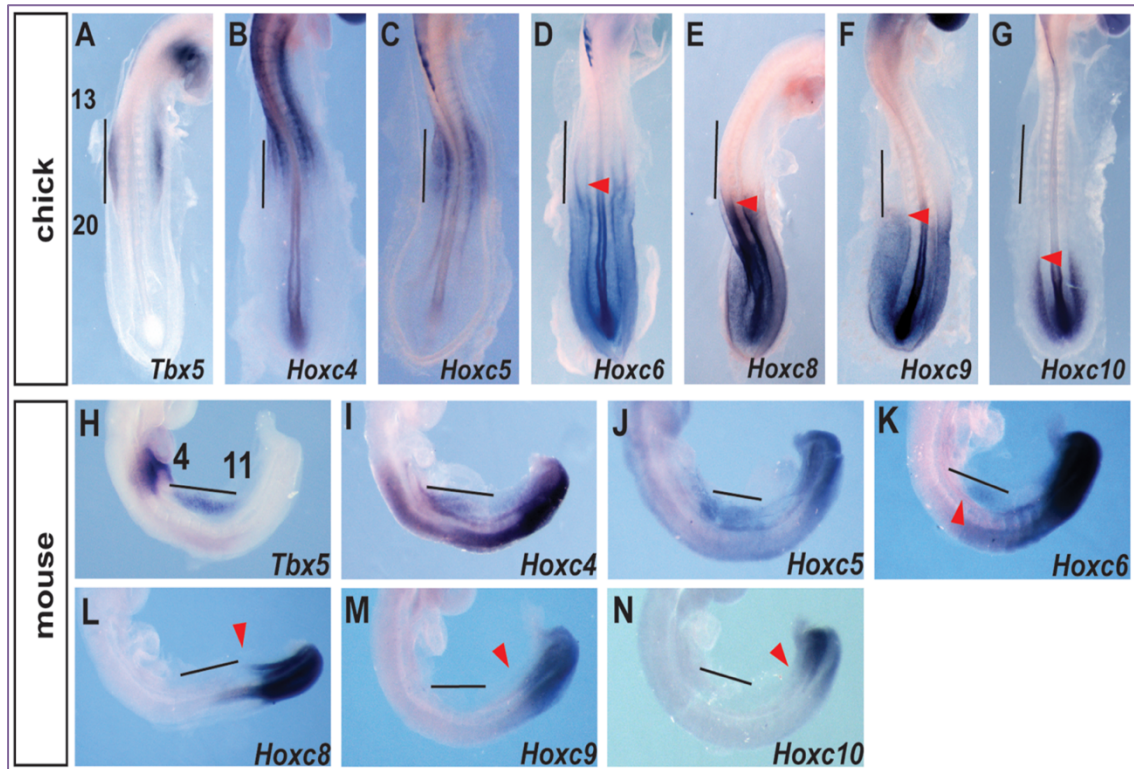


Figure 1.11: Comparison of *Tbx5* and *HoxC* gene expression domains. A–G. *In situ* hybridization for *Tbx5* (A) or *HoxC* genes (B–G) on 20–22 somite stages chick embryos. *Tbx5* is expressed lateral to somites 13–20. The forelimb-forming region is indicated by the vertical bar. The red arrowheads indicate the rostral extent of expression. H–N. *In situ* hybridization for *Tbx5* (H) or *HoxC* (I–N) on 11–13 somite stages mouse embryos. *Tbx5* is expressed lateral to somites 4–11. The forelimb-forming region is indicated by the vertical bar. The red arrowheads indicate the rostral extent of expression. Images taken from (Nishimoto *et al.*, 2014).

1.4.7. Hox gene regulation.

Mechanisms of somite formation and patterning are largely conserved along the length of the body axis. Different segments acquire different axial identity through the action of Hox transcription factors (Chen *et al.*, 2013). How this axial identity then impacts different morphologies is not clear, although I propose that this is in part through the action of Hox genes. Identification and characterization of the molecular signatures responsible for the functional specificity of Hox genes and other differentially expressed genes identified will help in our understanding of the patterning activity of these proteins (Casaca *et al.*, 2016).

For example, Hox paralogous group 6 genes, the rib-promoting genes, have been studied extensively and are known to control rib formation in part by regulating an enhancer, coined

the H1 enhancer, that controls *Myf5/6* expression in the mesoderm (Casaca *et al.*, 2016). However, how these proteins do this is only partially understood.

Hoxb6 has three conserved regions important for its activity - the linker region (LR), the homeodomain (HD) and the hexapeptide (HX). The linker region in Hoxb6, connecting the homeodomain and the hexapeptide, is essential for hoxb6 rib-promoting activity by bringing additional regulatory factors, such as Pax3, together to target DNA elements. In other words, the LR provides specific activities to the Hoxb6 protein through a Hox- Pax3-H1 enhancer complex. However, it was found, through loss of function mutations, and alongside subsequent experiments involving the substitution of the Hoxb6 LR with that of a gene lacking rib-inducing properties, that it is not sufficient to confer rib-inducing properties to other Hox proteins, but it is required (Casaca *et al.*, 2016). This suggests that the rib-inducing properties of Hoxb6 require the presence of additional characteristics that are specific to this protein.

But the LR is not involved in the formation of the Hox-Pax3-H1 enhancer complex, as was discovered after a Hoxb6 mutation lacking the LR still produced the complex but was unable to produce ribs. Instead, it recruits additional factors required to activate functional properties of the complex. It is thought that the HX or HD regions either side of the LR make distinct contributions to the formation of the complex and that regulatory interactions among these regions are required to properly control protein activity (Casaca *et al.*, 2016).

Understanding these interactions and how they trigger Hox functional activity requires the identification and characterisation of the factors that interact with these protein regions. With this project, some of the additional factors that confer specific functional activity to Hox paralogues may be uncovered. In future, a better understanding of their interactions will add to complete the story of different vertebrae morphologies, including rib development.

1.5. Enhancers

There are different mechanisms employed to regulate gene expression patterns. Some of these mechanisms are gene specific and others regulate more than one gene at a time. Much of the human genome consists of regions that do not code for any proteins. These noncoding regions were described as 'Junk DNA' - DNA regions with no purpose. But, in 1961 a study unveiled the regulatory control mechanism behind the lac operon within noncoding regions of DNA (Jacob and Monod, 1961). Distal cell specific regions of open chromatin were identified far from any expressed genes. It was shown that these regions 'enhanced' gene expression in *cis*, and the

term enhancer was first used (Banerji *et al.*, 1981). Many more have worked to characterise these *cis*-regulatory elements (CRE) since (Doane and Elemento, 2017). CREs are typically non-coding regions of DNA which contain binding sites for transcription factors (Wittkopp and Kalay, 2011). Promoters and enhancers are the best understood, but CREs also cover silencers and operators.

It is currently believed that the human genome contains roughly 500,000 enhancer regions, which are used to regulate gene expression during development and homeostasis (Gasperini *et al.*, 2020). However, despite the obvious importance of enhancers to both disease biology and fundamental biology, far more work is needed to fully characterise and understand the functional importance of these regions.

Enhancers are defined as short sequences of DNA that contain transcription factor binding motifs that increase the likelihood of transcription of one or more genes through a *cis* regulatory mechanism (Ong and Corces, 2011). It was initially thought that enhancers worked by direct interactions with RNA polymerase II, yet now it is believed that enhancers work by an indirect interaction (Kolovos *et al.*, 2012). They can be located many hundreds or even thousands of base pairs away from its promoter defined as the site where RNA polymerase binds to initiate transcription and can be located either up or down stream of its targets. To increase the likelihood of transcription of a gene, the enhancer must interact with its respective promoter, which could be located many base pairs away. It is the looping capability of DNA that allows the activators (such as transcription factors) bound to enhancer sequences to encounter the RNA polymerase or the general transcription machinery bound to the promoter and increase rate of transcription (Krivega and Dean, 2012). It is currently thought that DNA binding TFs bound at both the promoter and the enhancer recruit 'looping factors' that help to form the loop (Krivega and Dean, 2012). Fig 1.12 gives a diagrammatic representation of this looping behaviour.

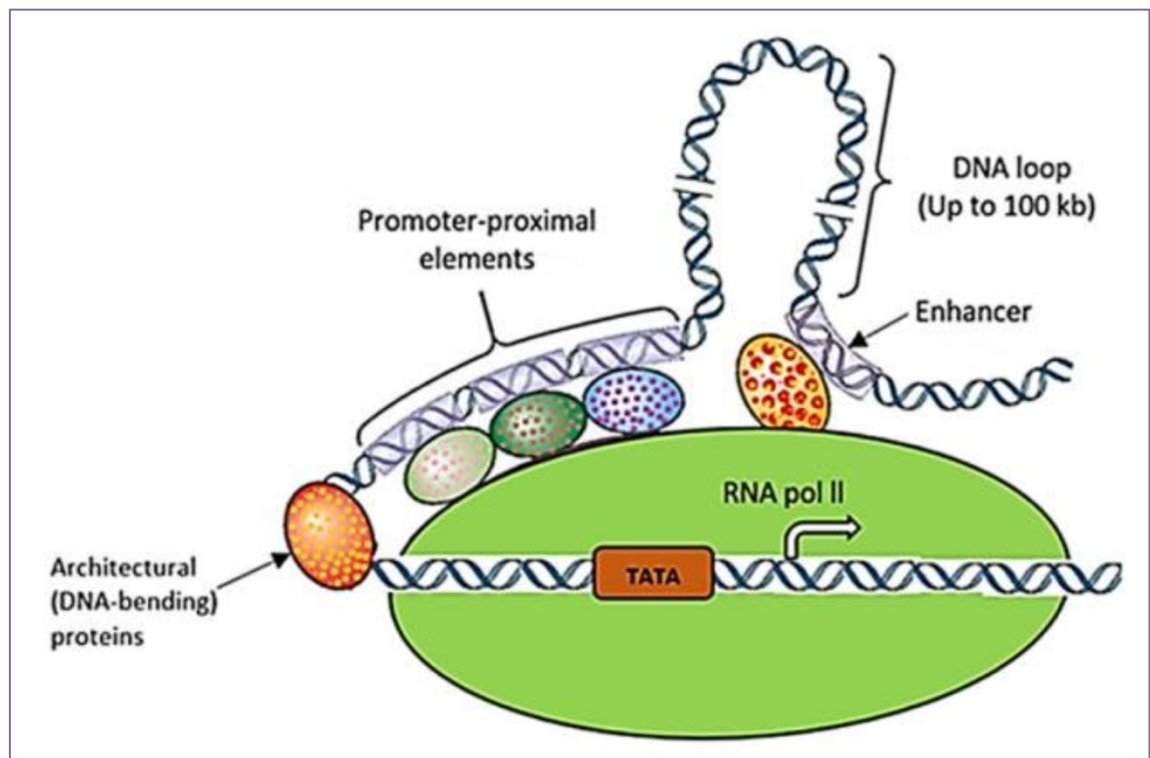


Figure 1.12: Enhancer mode of action theory. Image from (Li *et al.*, 2015).

There are a few distinguishable characteristics of enhancer regions. They are usually linked with certain properties of chromatin, and active enhancers are typically bound by transcription factors. As enhancers work spatiotemporally to regulate gene expression, they are found in different states depending on chromatin accessibility. Developmental enhancers undergo progressive changes that depend on TF activity and chromatin remodelling (Peng and Zhang, 2018).

Many mammalian genes have now been shown to be under the control of more than one single enhancer and, in most cases, enhancers tend to be cell specific, but it is not always the case. This gives rise to complex patterns of gene expression (Perry *et al.*, 2011) and can result from the additive action of different enhancers with cell type or tissue specific activities and gene expression (Barolo, 2011). There are 'primary' and 'secondary' enhancer regions, with the secondary enhancer often being described as a shadow enhancer (Barolo, 2011). It is currently thought that the presence of two enhancers with overlapping activity can provide a robustness to gene expression, establishing precise boundaries of gene expression (Frankel *et al.*, 2010). For example, in mouse embryos, *Shh* transcription is controlled by two enhancers. When activated by the two distinct enhancers, expression is restricted and robust in the floor plate. However, transcription in the limb bud is controlled by only one enhancer and gives a more sporadic expression pattern (Amano *et al.*, 2009).

The sequential activation of vertebrate Hox genes appears important to enable these genes to function correctly in embryogenesis and limb development and a selective pressure on the vertebrate, but not fly Hox genes appears to maintain their clustered organization. Shared enhancers apparently play important regulatory roles in the activation and maintenance of Hox gene expression in the vertebrate complexes (Mann, 1997; Suemori *et al.*, 2000)

1.6. *Project Aims and Objectives*

1.6.1. Aims

Sequencing-based approaches allow a more quantitatively accurate result through counting the number of transcripts, compared to spot intensities in microarrays, which fall in non-discrete gradients. An initial mechanism I would like to carry out is to characterise the molecular profile of axial identity across the cervical-thoracic transition using a variety of bulk sequencing methods. Any changes observed may indicate possible mechanisms and/or processes involved in defining the C-T transition. Specifically, I would like to remove six somites, three cervical and three thoracic, across this axial boundary and detect patterns in the molecular profile of the transitioning somites. I identified regions of open chromatin using ATAC-Seq and then associated these accessible chromatin regions with expressed genes using RNA-Seq transcriptomics.

This project aims to establish molecular signatures and cell lineages across the cervical-thoracic axial levels using developing somites of vertebrate embryos. I investigated how these discrete but highly related mesoderm lineages acquire axial level specific identity in early somites by characterising the molecular profiles of progenitors. I determined the accessible chromatin and transcriptome signatures of somites using ATACseq and RNAseq. I will focus on axial positions that produce prospective cervical and thoracic somites, with anatomically discrete features such as ribs. In depth analysis of these data aimed to uncover potential *cis*-regulatory elements and associated gene expression changes that correlate with positional identity along the cervical-thoracic boundary, as well as musculoskeletal cell fate acquisition. Experimental embryology combined with genomics analyses is used to validate and characterise putative enhancer elements. Overall, this is project used genomics approaches to investigate the molecular mechanisms that underpin a classic embryology experiment with the aim to provide novel mechanistic insights.

1.6.2. Objectives

1. To characterise the molecular profiles of somites with different axial identities by sequencing somites across the CT-boundary.
 - A. To determine accessible chromatin landscapes of prospective cervical and thoracic somites by using ATAC-Seq.
 - B. To determine the dynamics of gene expression changes by using RNA-Seq.
 - C. To examine the changes in these molecular signatures across the boundary of these two domains by performing a differential analysis.
2. To validate candidate CREs involved in the regulation of genes important for axial level specific lineages.
 - A. Candidate *cis*-regulatory elements, which may target Hox genes critical for mediating C-T transition (Hoxc5, Hoxc6, Hoxc8), will be identified and experimentally validated.
 - B. Investigate processes that are potentially regulated by Hox genes in somites and their derivatives.
 - C. To address how Hox genes interact with somite patterning pathways and cell differentiation genes to regulate vertebral morphologies.

CHAPTER 2: MATERIALS AND METHODS

2.1. Embryo development and preparation

2.1.1. Culture of chicken eggs

Fertilised white leghorn chicken eggs were ordered weekly from Henry Stewart & Co. Ltd, UK and stored at 17°C prior to incubation. To initiate development, the eggs were then transferred to a humid incubator and kept at 38°C for the length of time necessary for the required Hamburger and Hamilton (HH) stage. Stage HH12 (~33 hours) and stage HH14 embryos (~48 hours) were used for RNAseq, ATACseq and for *in ovo* injection and electroporation. Stage HH4 (~18 hours) were used for early EC culture injection and electroporation.

2.1.2. Embryo harvesting

2.1.2.1. For in situ hybridisation

Embryos were harvested from eggs by cutting away connecting tissues with fine scissors and removing the embryos using a forceps. The embryos were then placed in a Petri dish containing 1X PBS (Phosphate-buffered saline) and dissected free of extraembryonic membranes and other attached tissues in order to increase visibility of markers for WISH. The embryos were fixed in 4% paraformaldehyde (PFA)(SIGMA; P6148)/DEPC(Diethyl pyrocarbonate)-PBS at 4°C overnight.

2.1.2.2. For ex ovo injection and electroporation

Stage HH4 embryos were explanted together with the vitelline membrane and anchored to a filter paper carrier under tension for early chick (EC) culture (Chapman *et al.*, 2001). Embryos were transferred ventral side up onto a 30mm petri dish containing a 2.5ml substratum of agar-albumen and incubated at 38°C for 19-26 hours.

2.1.2.3. For RNA extraction

Embryos were dissected as previously described (section 2.1.2.1) but instead immediately collected in ice-cold DEPC-treated PBS and snap frozen in liquid nitrogen. Frozen samples were stored at -80°C until processed for RNA extraction (see below).

2.1.3. Embryo dehydration

Fixed embryos were dehydrated in a series of MeOH(methanol)/1X PBST(PBS with 0.1% Tween-20) dilutions by washing first in 25% MeOH/1X PBST, then 50%, then 75%, and finally 100%

methanol. Individual washes were carried out for ~15 min each. Once dehydrated, embryos were then stored in fresh methanol at -20°C for a maximum of three months.

2.1.4. Embryo rehydration

Before proceeding with the whole mount *in situ* hybridisation protocol, dehydrated embryos were rehydrated by washing them in 75% methanol, then 50% and finally 25% methanol/PBST. Individual washes were carried out for ~15 min each.

2.2. RNA probe synthesis by PCR

2.2.1. RNA extraction

Snap frozen stage HH14 embryos were thawed on ice and centrifuged for 2 min at 5000 rpm before removal of excess 1X PBS supernatant. RNA extraction was then carried out following manufacturer's instruction from the Qiagen kit, which included a protein digestion, lysis with RNA Lysis Buffer, gDNA removal via Spin-Away™ Filter, ethanol precipitation, column binding, DNase on the column digest, column preparation with two washes, and finally elution from the column. The RNA was eluted using 20µl of RNase free water before immediately stored on ice. The concentration and quality of RNA samples were verified using a NanoDrop spectrophotometer with an expected ratio of over 1.8 for readings at 260/280 and 260/230nm, indicating that the sample was free of residual ethanol, which inhibits enzymes in relevant downstream reactions.

2.2.2. cDNA synthesis

To generate gene specific probes, RNA was reverse transcribed to cDNA using Superscript®II Reverse Transcriptase (Life Technologies; 18064-022). To do so, 1µg of RNA extracted from stage HH14 chick embryos (see above) was added to 1µl oligodTs (0.5 µg/µl) and 1µl dNTP mix (10mM) (Promega). The volume was made up to 12µl using sterile RNase free water (SIGMA; W4502) and the mix was incubated for 5 min at 65°C. After incubation and a quick chill on ice, the following solution was added:

Table 2.1: cDNA synthesis reaction setup

Reagent	Volume (µl)
5x First Strand Buffer	4
0.1M DTT (Promega; P1171)	2
RNasin® (Promega; N2611)	1

The combined mix was incubated for 2 min at 42°C before adding 1µl of the Superscript II Reverse transcriptase. The entire solution was incubated for a further 50 min at 42°C and then inactivated at 72°C for 15 min. The cDNA template was then stored at -20°C until needed.

2.2.3. DNA template preparation

2.2.3.1. Amplification of cDNA

A standard Polymerase Chain Reaction (PCR) was performed to amplify 500-1000bp fragments of coding sequences of genes of interest from cDNA (Table 2.1) using the below PCR mixes (Table 2.2 & Table 2.3). Each pair of gene specific primers used (Appendix 1) were designed to have little to no complementary to other genes or gene variants using NCBI Primer Blast.

Table 2.2: cDNA amplification reaction setup

Reagent	Volume (µl)
cDNA	1
5X KAPA Long Range buffer (no Mg)	5
MgCl ₂ (25mM)	1.5
dNTP mix (10mM) (Promega)	0.5
KAPA Long Range polymerase (2.5 U/µl)	0.25
Forward primer (20µM)	1.25
Reverse primer (20µM)	1.25
H ₂ O (SIGMA; W5402)	14.25

The reaction conditions were as follows:

Table 2.3: cDNA amplification PCR conditions

PCR step	Temperature (°C)	Time	Cycles
Enzyme activation	94	4 min	1
Denaturing	94	30 sec	35
Annealing	55	30 sec	
Extension	72	2 min	
Final extension	72	7 min	1

5µl of each sample was then loaded onto a 1% agarose TAE gel in order to establish success of the PCR reaction. Molecular weight markers (1kb and 100bp DNA ladder; NEB) were used to establish product size. Following successful amplification of the gene of interest, the PCR products were purified following manufacturer's instructions using the PCR Purification kit (Qiagen; 28106). Briefly, this involves the addition of 5:1 volume of Binding Buffer to the remaining PCR product, followed by two washes in Wash Buffer and a final elution from the column with 20µl Elution Buffer.

2.2.3.2. pGEM-T-Easy cloning

Purified PCR products were then ligated into pGEMT-Easy Vector System I (Promega; A1360). This vector was selected as it contains both M13 forward and reverse primer binding sites, as well as SP6 and T7 RNA polymerase promoters, flanking the insertion site. As well as this, the multiple cloning site (MCS) within the pGEM-T Easy Vector, where the gene of interest is ligated, is located within the lacZ gene, thus giving us an indication of a successful ligation reaction when the absence of active B-galactosidase leads to the growth of white colonies instead of blue on LB plates containing X-gal. The cloning strategy was as follows:

Table 2.4: pGEM ligation reaction setup

Reagent	Volume (µl)
pGEM®-T Easy Vector (25ng/µl)	0.5
T4 DNA Ligase	0.5
2X Rapid Ligation Buffer	5
Purified PCR Product	4

After mixing the components, the reaction was incubated at 4°C overnight before transforming into competent cells.

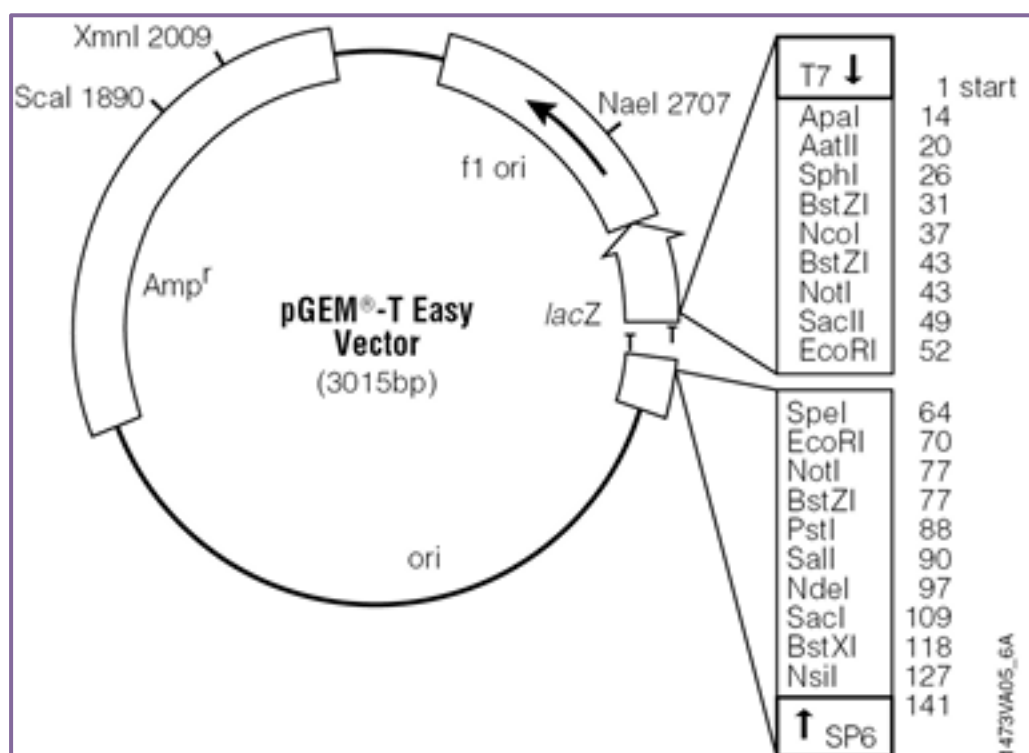


Figure 2.1: pGEM-T Easy vector map. The plasmid contains an Ampicillin resistance cassette (Amp^r) and the multiple cloning site within the lacZ gene, enabling for blue/white colony selection in Escherichia coli (E. coli) cells.

2.2.3.3. Plasmid transformation

5µl of the ligation mixture for transformation was added to 50µl DH5-alpha competent Escherichia coli cells in a 1.5ml Eppendorf tube and incubated on ice for 30 min. Cells were then heat-shocked at 42°C for 90 seconds to allow plasmid uptake by the bacteria. Following heat-shock the cells were then immediately incubated on ice for 3 min. To allow plasmid expression, 1ml of LB-media was added to the DH5-alpha bacteria and incubated at 37°C for 1hr 30 min with constant shaking. After incubation, the cells were pelleted at 6000g for 5 min, 900µl supernatant was discarded and the cells were resuspended in the remaining 100µl supernatant. The resuspended cells were then plated on selective LB-agar plates (Carbenicillin 100µg/µl) coated with 50µl X-gal (20mg/ml) for future blue-white colony selection and incubated overnight at 37°C.

2.2.3.4. Colony PCR

White colonies were selected from each plate and colony PCR reactions were performed to identify positive colonies with an insert of the expected product size. Each colony was picked using a sterile P20 tip, streaked onto a fresh bacterial plate, incubated overnight at 37°C, and then dipped into a PCR tube containing 10µl of the following PCR mix:

Table 2.5: colony PCR reaction setup

Reagent	Volume (μl)
2X BioMix Red Buffer (Bioline)	5
Forward M13 primer (20μM)	1
Reverse M13 primer (20μM)	1
H ₂ O (SIGMA; W5402)	3

The reaction conditions were as follows:

Table 2.6: colony PCR conditions

PCR step	Temperature (°C)	Time	Cycles
Enzyme activation	94	4 min	1
Denaturing	94	30 sec	35
Annealing	55	30 sec	
Extension	72	2 min	
Final extension	72	7 min	1

Colonies containing the correct product size were then transferred into culture tubes of 5ml LB medium and incubated at 37°C overnight before plasmid mini preparations were performed according to manufacturer's instructions (Qiagen; SBS27106). In short, this involved pelleting the bacteria before resuspending in Resuspension Buffer, mixing in Lysis Buffer and incubating in Neutralization Buffer. Following centrifuge in a spin column, the column was washed twice with Wash Buffer before elution with 20μl Elution Buffer. Sequencing was verified using either the SP6 or the T7 primer. This also determined the orientation of the inserted cDNA fragment.

2.2.4. Probe synthesis

Labelled anti-sense RNA probes were generated from the amplified DNA template using the reagents in the table below (Table 2.7). To make a linear DNA template for probe synthesis, with the correct orientation (5'-3'), appropriate sets of primers were selected to amplify by PCR the inserted gene fragments.

Table 2.7: probe synthesis reaction setup

Reagent	Volume (μl)
H ₂ O (SIGMA; W5402)	20
5X Transcription buffer (Promega)	10
DTT (100mM; Invitrogen)	5
RNasin (40 units/μl)	1
10X DIG-UTP Labelling mix (10mM; Roche)	5
T7 or SP6 RNA polymerase (20 units/μl) (Promega)	2
Template DNA	7

Transcription reactions were incubated for 16hr at 37°C for T7 polymerases and 40°C for SP6 polymerases before adding 1μl DNase (2 units/μl; Ambion) and incubating for 30 min at 37°C to degrade DNA templates. Probe integrity was examined by running 1μl on a 1% agarose TAE gel.

2.2.5. Probe purification

Probes were purified using Illustra ProbeQuant G-50 Micro Columns (GE Healthcare) following manufacturer's instructions. Finally, 50μl of probe was added to 10ml of Hybridisation Buffer and then stored at -20°C until use.

2.3. Whole mount *in situ* hybridisation (WISH)

2.3.1. Day 1: Hybridisation

Rehydrated embryos underwent two 5 min PBST washes before treatment with 15μg/ml Proteinase-K in PBST for 30 min. After treatment, embryos were rinsed in PBST and post-fixed in 4% PFA for 20 min. Again, embryos were rinsed in PBST in preparation for hybridisation. To do so, they were washed in a 1:1 PBST/Hyb Buffer solution for 5 min. After settling, the embryos were transferred into Hyb buffer solution and incubated for >2 hours at 65°C. After pre-hybridisation, the Hyb buffer solution was replaced with the Hyb buffer containing the probe and incubated overnight at 65°C.

- **Hyb-Buffer:** 50% v/v Formamide, 1.3x SSC (pH 5.0), 5mM EDTA (pH 8.0), 50μg Yeast RNA (Sigma R6750), 0.2% v/v Tween-20, 0.5% w/v CHAPS, 100μg/ml Heparin (Sigma, H3149) in DEPC-H₂O.
- **Wash-Buffer:** 50% v/v Formamide, 1.3x SSC (pH 5.0), 0.2% v/v Tween-20 in DEPC-H₂O.

2.3.2. Day 2: Immunodetection

The probe solution was removed and stored at -20°C for reuse. The embryos were then rinsed twice in Hyb buffer and washed 4 times in wash buffer for 30 min at 65°C. Following this, embryos were washed in a 1:1 Wash buffer/MABT solution at 65°C. Subsequently, 3 rinses in MABT followed and two more 30 min washes in MABT at room temperature. Embryos were then pre-blocked in block solution (MABT/2% Roche blocking reagent; 11096176001) for 1hr at room temperature and then incubated in block solution with 10% goat serum for 2 hr at room temperature. Finally, the solution was replaced with fresh block solution with 10% Goat serum including the antibody in a 1:2000 dilution (anti-Dig-AP Fab fragment; Roche; 11093274910) and incubated at 4°C overnight.

- **MABT:** 100mM Maleic acid (Sigma M0375), 150mM NaCl, 0.1% v/v Tween-20 (pH 7.5).

2.3.3. Day 3: Colour development

The antibody solution was removed, and embryos were washed in MABT 3 times for 1 hr. Following the washes, the pH was adjusted for colour development by two 10 min washes in NMTT (pH 9.5). Subsequently, the colour development solution (NBT/BCIP in NMTT, 7/9µl/ml respectively) was applied and incubated until a satisfactory signal was developed. Embryos were then transferred to 5X TBST for background stain removal. Colour reactions were occasionally paused overnight at 4°C in PBST and repeated the following day starting with NMTT.

- **TBST:** 20mM Tris, 150mM NaCl, 0.1% v/v Tween-20 (pH 7.5).
- **NBT:** 4-Nitro blue tetrazolium chloride (Roche 11087479001), stock at 75mg/ml in 70% v/v dimethylformamide. Aliquot and store at -20°C. Protect from light.
- **BCIP:** 5-bromo-4-chloro-3-indolyl-phosphate, 4-toluidine salt (Roche 10760994001), stock at 50mg/ml in dimethylformamide. Aliquot and store at -20°C. Protect from light.
- **NMTT:** 100mM NaCl, 100mM Tris-HCl (pH 9.5), 50mM MgCl₂, 1% v/v Tween-20 in DEPC-H₂O.

2.4. *Histological sectioning*

Fixed embryos were transferred to a 20% sucrose (Fisher Scientific; S5-3)/PBS solution overnight at 4°C prior to cryopreservation. Tissues were then embedded in moulds containing OCT Embedding Matrix (Cellpath™, Fisher Scientific; 1521277) prior to sectioning and frozen on a flat surface of dry ice for 30 min before storage at -20°C. The embedded samples were attached to a mounting plate with OCT and placed on a -50°C cooler plate in a cryostat (LEICA CM 1959) for 20 min until secured. Subsequently, the sample was mounted to the -21°C cryostat and aligned to the blade. Sections were first trimmed in 50µm slices until the tissue was exposed. The

cryostat was then switched to fine mode (20µm slices) and sections were cut and transferred to Superfrost-Plus slides (ThermoScientific; J1800AMNZ).

The finished slides were collected and dried at 4°C overnight. The slides were then washed twice in 1X PBS for 5 min and mounted in Hydromount (National Diagnostics). The slides were dried again overnight at 4°C before imaging using AxioVision software on an upright microscope (Zeiss).

2.5. In ovo grafting

2.5.1. Part 1: donor embryo

With a small, sharp scissors, the area around a Stage HH14 embryo was cut and the embryo transferred ventral side up to an agar dish using a spatula/spoon. All liquid was removed from the embryo to allow sticking to the agar dish without pinning.

With a Tungsten needle, very shallow cuts were made laterally along the length of PSM tissue equivalent to three somites. First the tissue was cut medially alongside the notochord and then laterally. One drop of Dispase solution (1.5mg/ml in DMEM/10mM Hepes pH7.5) was added to the embryo and then diluted with two drops of PBS. Nile Blue stain (prepared the day of the experiment) was added to the tissue by dragging a Nile Blue pen along the surface of the liquid above the embryo, in order to visualise the small lateral cuts made in the endoderm. Immediate removal of the endodermal layer was carried out by pulling the tissue away and to the side with the Tungsten needle. Once the endoderm was removed, the tissue was stained once more with Nile Blue. This will also help with orientation after the tissue is removed, as only the ventral side of the tissue will be stained. The aorta can be seen as a small white line along the edge of the somitic and presomitic tissue (opposite to the notochord). This was carefully pulled away using a Tungsten needle – if the aorta was cut, you could see a lot of blue ‘debris’, which are red blood cells that have been stained by the Nile Blue. Then remove the desired tissue by cutting along the outside of the tissue. Do not cut at the inside, but instead pull aside the notochord from the tissue. Following these lateral cuts, small incisions were made anteriorly and then posteriorly. At the posterior cut, a slanted incision was made to help with orientation in the host. Following these cuts, the Tungsten needle was slid beneath the tissue, allowing it to lift from the endoderm beneath.

Once removed, a small, pulled capillary needle at the end of a mouth pipette was used to remove the donor tissue. This was placed into a small petri dish with bovine serum albumin (BSA) to stop the Dispase from acting. The tissue can be left in this solution and placed on ice for a few hours before insertion into the host at RT.

2.5.2. Part 2: host embryo

Stage HH10 eggs were incubated with the blunt end up where they were then punctured with a forceps and a large hole above the embryo was created for easy accessibility. A solution of 1:500 PBS/Pen Strep (PS) with black India Ink (Winsor & Newton) was injected beneath the embryo using a 1 ml syringe with a 25Gx 5/8" needle for contrast and then 2 drops of 1:500 PBS/PS-solution was applied above the embryo to increase elasticity of the vitelline membrane. After removal of all shell membranes covering the embryo, a small incision was made in the vitelline membrane above the graft site of the embryo using the tip of a 25Gx 5/8" syringe needle.

Using a Tungsten needle, very shallow and small cuts were made on the medial side of the tissue to be removed. Following the shallow cuts, the tissue was stained with Nile Blue to visualise the ectoderm. The top ectodermal layer was then pulled away from the tissue with the Tungsten needle and the underlying tissue was stained again with Nile Blue. With a pulled capillary needle at the end of a mouth pipette, all tissue from the graft area was removed. When unsure that all tissue had been removed, staining with Nile Blue was repeated. Once the tissue had been removed, the donor tissue was placed into the space using a pulled capillary needle at the end of a mouth pipette. Using a blunted glass pasteur pipette that had been melted into a round ball shape over a Bunsen burner, the donor tissue was pushed into place. With the Tungsten needle, the host ectoderm was pulled over the donated tissue and pinned in place (Figure 2.2).

Following grafting, ~3ml of thin albumin was removed from the eggshell using a 2ml syringe with a 21Gx 5/8" needle to ensure the embryo does not touch the sealing tape. The eggshell was finally resealed with tape and immediately returned to the incubator at 38°C for ~2 days. Grafted embryos were then harvested and fixed for WISH (see section 2.1.2.1).

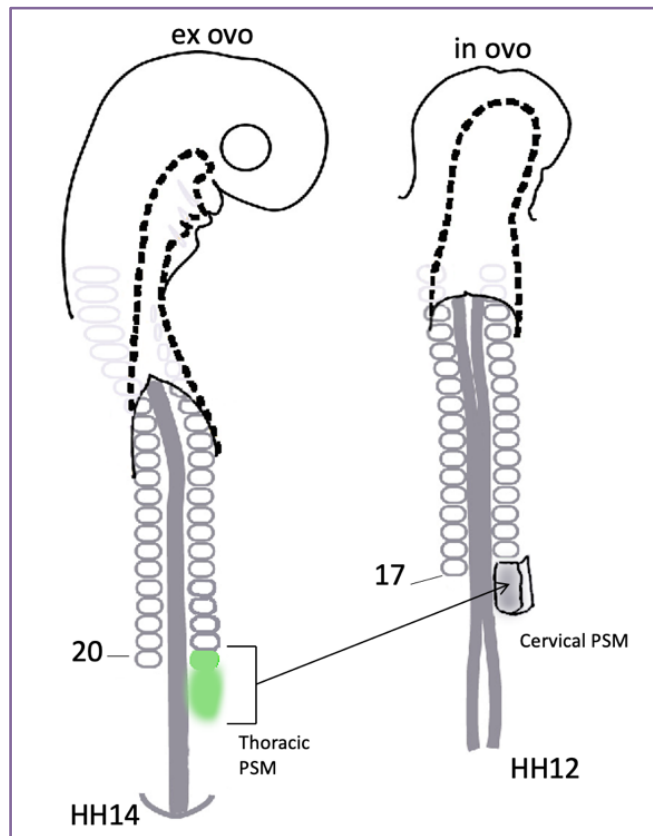


Figure 2.2: Schematic of heterotopic graft. Here the three anterior-most prospective thoracic somites (numbered 20-22) are removed from a donor stage HH14 and then inserted *in ovo* in place of the three posterior-most cervical somites (numbered 17-19) in a host stage HH14.

2.6. RNAseq

2.6.1. Somite dissections

Eggs were incubated for 52hr or 56hr, until they reached stages HH13 and HH14+. Specifically, embryos used had developed between 19 and 22 pairs of somites. The embryos were removed from the egg and placed into a petri dish with phosphate-buffered saline (PBS), where they were rinsed free of contaminating debris and the vitelline membrane was removed. The embryos were then pinned down, dorsal side facing upwards, onto a Sylgard-coated dish with 6 insect pins.

150ul of Dispase solution (1.5mg/ml in DMEM/10mM Hepes pH7.5) was added to the embryo for 7 min, followed by 150ul of Trypsin solution (0.05% Trypsin/0.53mM EDTA in HBSS, with Sodium Bicarbonate, without Calcium and Magnesium) for a further 7 min once Dispase action was complete. 150ul BSA was then added to the embryo and somite removal was initiated.

The removal of the somites involved cutting the ectoderm covering the somites with a tungsten microneedle and peeling it away. The boundaries of the somites were then cut away from the adjacent structures using the same microneedle and the entire somite was pulled out of place using a scooping motion with the needle (see Figure 2.3).

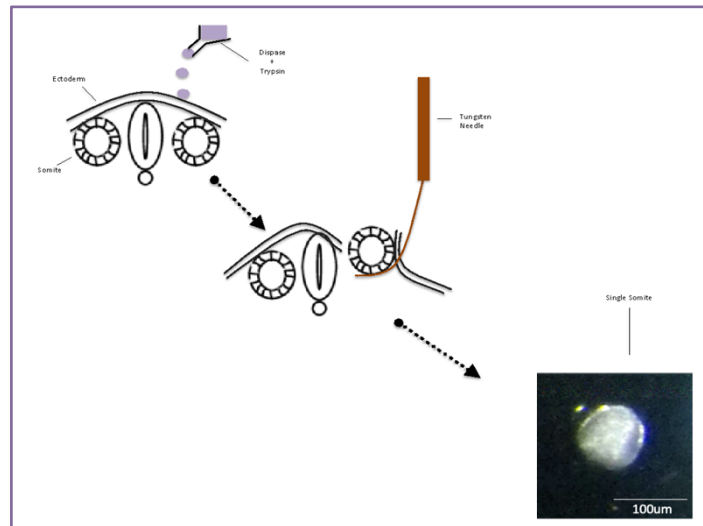


Figure 2.3: Schematic of somite isolation procedure. (A) Dispase and trypsin is added to the ectoderm of an embryo pinned to a Sylgard dish. (B) The ectoderm is then peeled back, and the somite removed with a Tungsten needle. (C) A singular isolated somite.

For stage HH14 (22 somite pairs) the six most recently formed consecutive somites across the cervical-thoracic boundary on the right side of the embryo were removed (somites no. 17-22) and collected individually. For stage HH13 embryos (19 somite pairs), the three most recently formed consecutive somites sitting just before the cervical-thoracic boundary on the right side of the embryo were removed (somites no. 17-19) (Figure 2.4). Four biological replicates were generated.

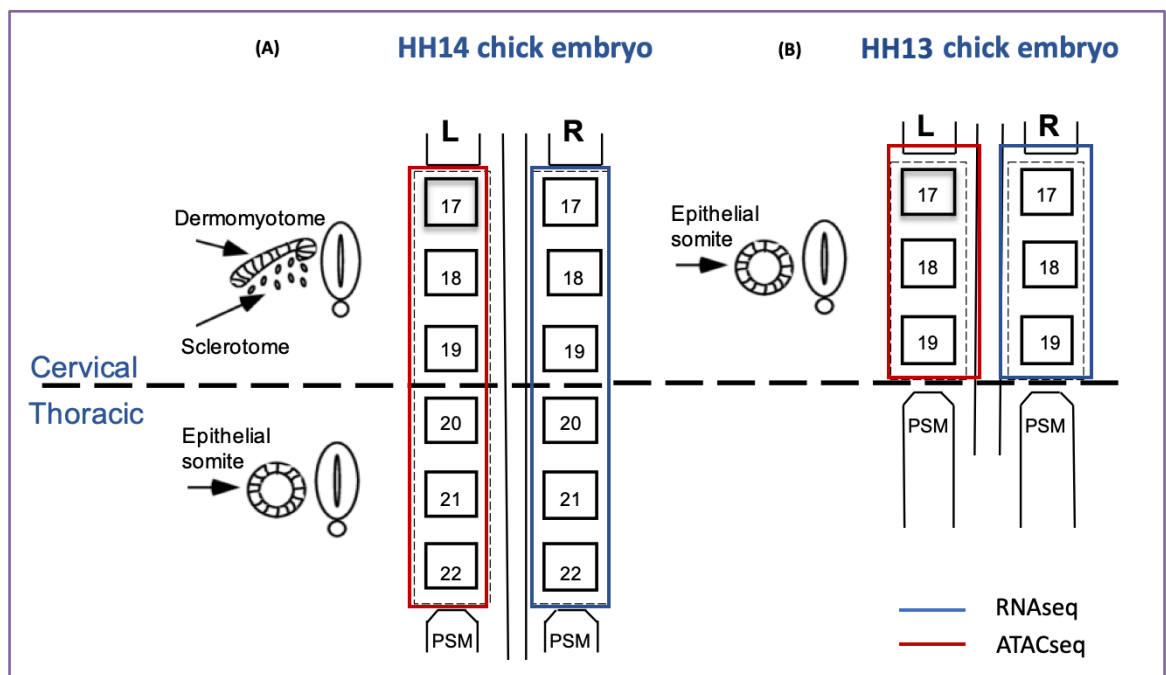


Figure 2.4: Schematic of RNAseq/ATACseq experimental setup. (A) 6 most recently formed somites (no.17-22) across the C-T boundary in HH14 chick embryo. (B) 3 most recently formed cervical somites (no. 17-19) in a HH13 chick embryo. Somites are depicted by a square shape with their corresponding number in the centre above a segment of presomitic mesoderm (PSM). Somites outlined in blue were removed from the right (R) for RNAseq, and somites outlined in red were removed from the left (L) for ATACseq. The maturation stage of each somite (epithelial vs. maturing somite) are to the left of each figure.

Individual somites were placed into separate wells of a 96-well plate containing 10µl RLT+ lysis buffer and repeatedly pipetted up and down to attain a single cell suspension. Samples were stored at -80°C (no more than 6 months) until used to produce RNA-Seq libraries.

2.6.2. Library preparation

Libraries were prepared and sequenced on the Illumina HiSeq4000 platform (150bp paired end) at the Earlham Institute using the Smart-seq2 protocol (Picelli *et al.*, 2014) for amplification of RNA from low concentrations.

First, Mastermix 1 was prepared on ice in a 1.5ml Eppendorf tube using the following reagents:

Table 2.8: Mastermix 1 reagents

Reagent	Volume per well (µl)
Oligo-dT ₃₀ VN primer (100µM)	0.1
dNTP mix (10mM)	1
Nuclease Free Water	0.9

Mastermix 2 was also prepared on ice in a 1.5ml Eppendorf tube using the following reagents:

Table 2.9: Mastermix 2 reagents

Reagent	Volume per well (μl)
Nuclease Free Water	0.29
TSO oligo (100μM)	0.1
MgCl ₂ (1M)	0.06
Betaine (5M)	2
DTT (100mM)	0.5
5X Superscript first strand buffer	2
RNAse Inhibitor (20U/μl)	0.25
Superscript II (200U/μl)	0.5

Using a dispensing pipette fitted with a 200l tip, 2μl of Mastermix 1 was dispensed into each sample. The plate was sealed and spun down for 10 sec at 500g. The plate was then incubated on a thermal cycler at 72°C for 3 min to anneal the oligodTs to the poly-A tail of the mRNA transcripts before immediate transfer of the plate to ice for 1 min. The plate was then centrifuged to collect any condensation.

5.5μl of Mastermix 2 was then deposited into each well and the plate was then sealed and spun down for 10 sec at 500g. The following thermocycling programme was then carried out on the PCR machine:

Table 2.10: Smart-seq2 PCR conditions

Temperature (°C)	Time (min)	Cycles
42	90	1
50	2	10
42	2	
70	15	1

As the reverse transcription neared completion, Mastermix 3 was prepared using the following reagents:

Table 2.11: Mastermix 3 reagents

Reagent	Volume per well (μl)
2X Kapa Hifi Hotstart Readymix	12.5
IS PCR primers (10μM)	0.25
Nuclease Free Water	2.25

The plate was then spun down for 10 sec at 500g to collect any condensation and 15μl of Mastermix 3 was dispensed into each well. The plate was sealed and spun again for 10 sec at 500g. The following thermocycling programme was then carried out on the PCR machine:

Table 2.12: Smart-seq2 PCR conditions 2

Temperature (°C)	Time	Cycles
98	3 min	1
98	20 sec	21
67	15 sec	
72	6 min	
72	5 min	

Once complete, the plate was once again spun for 10 sec at 500g and purified using Agencourt AMPure XP beads (Beckman Coulter) following manufacturer's instructions. Samples were assessed for quality using an Agilent 2100 Bioanalyser in order to check for the required result of cDNA fragments ranging from ~500 - ~2,500nt in length. Libraries were then sequenced using paired-end 150bp reads on the Illumina HiSeq4000 platform at Novogene UK.

2.6.3. Transcriptome analysis

Four biological replicates were used for analysis. Adaptors were removed from raw paired-end sequencing reads and trimmed for quality using Trim Galore! (v.0.5.0) using default parameters. QC was performed before and after read trimming using FastQC (v.0.11.6) and no data quality issues were identified after checking the resultant QC reports. Processed reads were mapped to galGal6 cDNA using kallisto (v.0.44.0). Resultant quantification files were collated to generate an expression matrix. Differential expression and pathway analyses were then conducted using the DESeq2 package in R Studio (v.1.3.959) and default settings within the iDEP (v.9.51) web interface. GO term analysis (Gene Ontology) was carried out using g:Profiler, a web server for functional profiling and interpretation of gene lists.

2.7. ATACseq

2.7.1. Somite dissections

Samples were dissected as stated above (section 2.6.1). Somites were transferred immediately into 10µl of cold lysis/tagmentation buffer in the wells of a 96-well plate for 15 min, with intermittent pipetting to attain a single cell suspension. The plate was then spun for 10 sec at 500g and tagmentation was carried out for 1 hr at 37°C on a shaking thermomixer. The reagents used to create the lysis/tagmentation buffer come from the Illumina Nextera DNA kit (FC-121-1030) and are as follows:

Table 2.13: ATAC lysis/tagmentation buffer reagents

Reagent	Volume per well (µl)
PBS	3.3
H ₂ O (SIGMA; W5402)	0.4
Dig (1:1) w/ H ₂ O	0.1
Tween-20	0.1
NP40	0.1
2X buffer	5
TnF enzyme	1

2.7.2. Library preparation

Tagmented DNA was purified following manufacturer's instructions using Qiagen MinElute kit (Cat no. 28004) and then amplified using NEB Next High-Fidelity 2X PCR Master Mix (Cat no. M0543S) and a universal AD1 forward primer with individual reverse primers per sample, to use as index sequences for sample identification after sequencing (Appendix 1). The PCR conditions are as follows:

Table 2.14: ATACseq PCR conditions

Temperature (°C)	Time	Cycles
72	5 min	1
98	30 sec	11
98	10 sec	
63	30 sec	
72	1 min	

Library preparation was complete after further clean up using Qiagen PCR MinElute kit (Cat no. 28004) and Beckman Coulter XP AMPpure beads (A63880). Tagmentation size was assessed using Agilent 2100 Bioanalyser. Libraries were quantified with Qubit 2.0 (Life Technologies) and sequenced using paired-end 150bp reads on the Illumina HiSeq4000 platform at Novogene UK.

2.7.3. Transcriptome analysis

Four biological replicates were used for analysis. The analysis described in the following section was not carried out by myself, but instead by an expert ATACseq bioinformatician in the Münsterberg lab, Leighton Folkes, who had worked with ATACseq data previously (Mok *et al.*, 2021). Adaptors were removed from raw paired-end sequencing reads and trimmed for quality using Trim Galore! (v.0.5.0) a wrapper tool around Cutadapt and FastQC. Default parameters were used. Quality control (QC) was performed before and after read trimming using FastQC (v.0.11.6) and no issues were highlighted from the QC process. Subsequent read alignment and post-alignment filtering was performed in concordance with the ENCODE project's "ATAC-seq Data Standards and Prototype Processing Pipeline" for replicated data (<https://www.encodeproject.org/atac-seq/>).

Reads were mapped to the chicken genome galGal6 assembly using bowtie2 (v.2.3.4.2). The resultant Sequence Alignment Map (SAM) files were compressed to the Binary Alignment Map (BAM) version on which SAMtools (v.1.9) was used to filter reads that were unmapped, mate unmapped, not primary alignment or failing platform quality checks. Reads mapped as proper pairs were retained. Multi-mapping reads were removed using the Python script `assign_multimappers` provided by ENCODE's processing pipeline and duplicate reads within the BAM files were tagged using Picard MarkDuplicates (v.2.18.12) (<http://broadinstitute.github.io/picard/>) and then filtered using SAMtools.

For each step, parameters detailed in the ENCODE pipeline were used. From the processed BAM files, coverage tracks in bigWig format were generated using deepTools bamCoverage (v 3.1.2) and peaks were called using MACS2 (v.2.1.1) (parameters -f BAMPE -g mm -B -nomodel -shift -100 -extsize 200). Coverage tracks and peaks (narrow peak format) were uploaded to the UCSC Genome Browser as custom tracks for ATAC-Seq data visualisation.

2.7.4. Footprinting analysis

For clarification, the analysis described in this section was carried out by Leighton Folkes. Analysis of ATAC-Seq for differential accessibility was carried out in R (v.3.5.1) using the DiffBind package (v.2.8.0) with default parameter settings. Differential accessibility across samples was calculated using the negative binomial distribution model implemented in DESeq2 (v1.4.5). Computational footprinting analysis was conducted across samples using HINT-ATAC which is part of the Regulatory Genomic Toolbox (v.0.12.3) also using default parameter settings and the galGal6 genome.

2.9. Preparation of putative enhancers

2.9.1. gDNA isolation

Chick genomic DNA (gDNA) was extracted from snap frozen stage HH14 chick embryos according to manufacturer's instructions using the Invitrogen Purelink gDNA extraction kit (Cat no. K1820-00). The concentration and quality of gDNA samples were verified using a NanoDrop spectrophotometer.

2.9.2. Plasmid constructs

A standard PCR was performed to amplify 500-1000bp fragments of putative enhancers from gDNA using the below PCR mix. Each pair of primers used (Appendix 1) were designed with specific sequence tails to enable cloning into the pTk-Citrine reporter vector.

Table 2.15: putative enhancer amplification reaction setup

Reagent	Volume (μl)
gDNA (200ng)	2
5X KAPA Long Range buffer (no Mg)	10
MgCl ₂ (25mM)	3.5
dNTP mix (10mM) (Promega)	1.5
KAPA Long Range polymerase (2.5 U/μl)	0.5
Forward primer (20μM)	2.5
Reverse primer (20μM)	2.5
H ₂ O (SIGMA; W5402)	27.5

The reaction conditions were as follows:

Table 2.16: putative enhancer amplification PCR conditions

Temperature (°C)	Time	Cycles
94	3 min	1
94	15 sec	10
55	15 sec	
68	3 min	
94	15 sec	35
63	15 sec	
68	3 min	
72	4 min	1

The PCR product was then purified using Qiagen PCR Cleanup kit (Cat no. 28104) and 5µl was run on a 1% agarose gel to confirm the expected size of the product. Samples were then prepared for T4-mediated ligation and BsmBI digestion using the following reagents:

Table 2.17: BsmBI reaction setup

Reagent	Volume (µl)
PCR product (75ng)	x
pTk-Citrine nanotag vector (75ng)	1
T4 DNA Ligase buffer	2
BsmBI	1
T4 DNA Ligase	1
H ₂ O (SIGMA; W5402) to final 20µl volume	x

The reaction conditions were as follows:

Table 2.18: BsmBI PCR conditions

Temperature (°C)	Time (min)	Cycles
37	2	25
16	5	
55	5	1
80	5	1

As a final “clean-up” step, the PCR product underwent a plasmid safe reaction, thereby selectively removing any contaminating linearised bacterial chromosomal DNA from the product. 10.25µl of the BsmBI reaction was added to 1.25µl of Plasmid safe buffer, 0.5µl ATP (25mM) and 0.5µl plasmid safe enzyme before incubating for 1 hour at 37°C and another 30 min at 70°C.

5µl of each sample were then directly transformed into 50µl DH5-alpha competent *Escherichia coli* cells (section 2.2.3.3) and the final resuspended cells were then plated on selective LB-agar plates (Carbenicillin 100µg/µl) coated with 50µl X-gal (20mg/ml) for subsequent blue-white colony selection and incubated overnight at 37°C. Following blue-white colony selection and another overnight culture of a single colony in 5ml LB medium, plasmid mini-preparations were performed according to manufacturer’s instructions (Qiagen; SBS27106) and verified with sequencing (Eurofins).

Upon confirmation of the correct putative enhancer sequence, plasmid maxi-preparations were performed according to manufacturer’s instructions using the EndoFree® Plasmid Maxi Kit (Qiagen; 12362). The plasmid was eluted using 30µl of elution buffer and the concentration and quality of samples were verified using a NanoDrop spectrophotometer.

2.9.3. Ex ovo plasmid injection and electroporation

Microinjection needles were prepared by pulling glass capillaries (1.0mm O.D. x 0.78mm; Harvard apparatus, UK) on a micropipette puller (P-97, Sutter Instrument Co., CA), with the following settings: P = 300, Heat = ramp test value, Pull = 150, Vel = 80, Time = 130. Needles were loaded using Microloader™ (Eppendorf) and mounted onto the microinjector. Stage HH4 embryos were transferred ventral side up into an electroporation chamber containing ~3ml of Howard Ringers Solution (HR) and a 2mm² platinum cathode below the embryo. Embryos were injected with candidate enhancer plasmid DNA (1µg/µl) in 10% injecting dye (Fast Green FCF, Sigma-Aldrich; Cat no./ID F7252) between the primitive streak and the epiblast. As a positive control to test for electroporation efficiency, embryos were co-injected with a chromatin marker plasmid, H2B-mCherry (1ug/ul) (Howe *et al.*, 2012). Electroporation was performed by placing a 2mm² platinum anode over the embryo and applying five pulses (5V for 50ms at 100ms intervals) using a square wave electroporator (Intracel). The embryo was then returned to its respective 30mm petri dish (section 2.1.2.2), incubated for 19-26 hours, and observed under a fluorescence stereomicroscope at stage HH8-14.

2.9.4. *In ovo* plasmid injection and electroporation

Microinjection needles were prepared as above (section 2.9.3). Eggs were incubated with the blunt end up where they were then punctured with a forceps and a large hole above the embryo was created for easy accessibility. A solution of 1:500 PBS/Pen Strep (PS) with India Ink (Winsor & Newton) was injected beneath the embryo using a 1 ml syringe with a 25Gx 5/8" needle for contrast and then 2 drops of 1:500 PBS/PS-solution was applied above the embryo to increase elasticity of the vitelline membrane. After removal of all shell membranes covering the embryo, a small incision was made in the vitelline membrane above the injection site of the embryo using the tip of a 25Gx 5/8" syringe needle. A micromanipulator was then used to direct the needle into the PSM and neural tube for injection.

Stage HH10 chick embryos were injected at multiple sites along the PSM and neural tube on the right side only. Electroporation was performed by placing a 2mm² platinum anode on the left side of the embryo and a 2mm² platinum cathode slightly elevated on the right side and applying five pulses (60V for 50ms at 100ms intervals) using a square wave electroporator. Following electroporation, ~3ml of thin albumin was removed from beneath the eggshell using a 2ml syringe with a 21Gx 5/8" needle. The egg was then resealed with tape and returned to incubate at 38°C for ~20 hours.

2.9.5 Enhancer mutagenesis

For mutagenesis of specific sites in enhancers we utilised FastCloning methodology (Li *et al.*, 2011), where the vector and insert are PCR amplified separately (Figure 2.5). Forward and reverse primers were designed to overlap and create a mutation/insert where the putative critical footprint of each enhancer was located (Appendix 1). Each primer pair was designed with a ~16-base overlap with the ends of the amplified enhancer vector. Primers were also designed to target Ampicillin; the antibiotic resistance gene insert of the enhancer vector (see section 2.9.2).

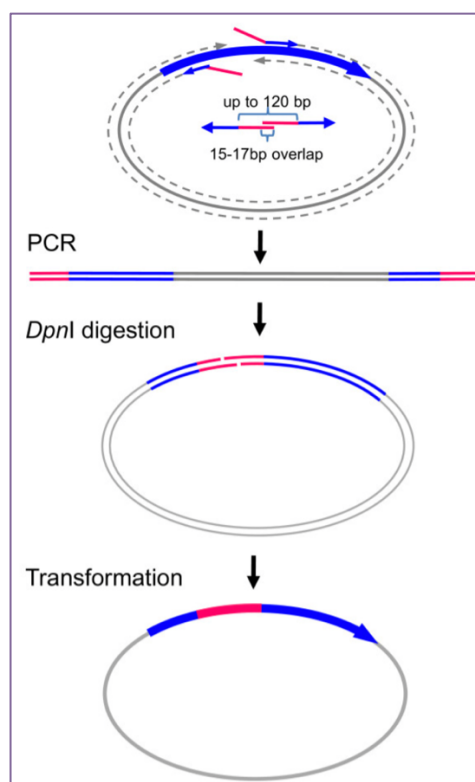


Figure 2.5: Mutagenesis PCR design. To replace a short stretch of DNA. The insert can be directly included in two primers for single PCR amplification of the cDNA along with the vector. The forward primer starts immediately downstream of the insertion site and has a tail with the 3' part of the insert. The reverse primer starts immediately upstream of the insertion site and has a tail with 5' part of the insert. Both primers only require a ~16-base overlap. Figure taken from (Li *et al.*, 2014).

Two standard PCR reactions were carried out per mutation needed. First containing the Forward primer for our mutation and the Reverse primer for Ampicillin. The second reaction then had the Forward primer for Ampicillin and the Reverse primer for our mutation. The reagents required for these reactions were as follows:

Table 2.19: Enhancer mutagenesis reagents

Reagent	Volume (μ l)
Plasmid (200ng)	1
5X buffer HF	5
DMSO	0.75
dNTP mix (10mM) (Promega)	0.5
KAPA Long Range polymerase (2.5 U/ μ l)	3.75
Forward primer (20 μ M)	1.25
Reverse primer (20 μ M)	1.25
H ₂ O (SIGMA; W5402)	15

The reaction conditions were as follows:

Table 2.20: Enhancer mutagenesis PCR conditions

Temperature (°C)	Time	Cycles
98	30 sec	1
98	10 sec	35
60	30 sec	
72	3 min	
72	5 min	1

Following confirmation of PCR products on a 1% agarose gel, the remaining unpurified PCR reactions, containing amplified vector and insert, are mixed and digested with 1µl *DpnI* for 1 hour at 37°C. The digested mixture is then transformed into chemically competent *E. coli* cells (see section 2.2.3.3). Following confirmation of the correct product through sequencing, the plasmids were then maxi-prepped following manufacturer's instructions (Qiagen; 12362) and electroporated *in ovo*.

2.10. HOXC6 and HOXC8 target validation

2.10.1. Identifying downstream targets of Hox gene

By combining data from the RNAseq dataset with differential footprints called from the ATACseq dataset, a selection of 10 putative downstream targets of HOXC6 and/or HOXC8 were identified (Appendix 1). Genes of interest were identified by selecting for those that had differentially called footprints as well as differential expression in the RNAseq dataset. A manual search for HOXC6 and HOXC8 footprints within 5kb upstream and downstream (or until the nearest gene) inside potential enhancer sites for the differential genes was then carried out. If found, those genes with a HOXC6/HOXC8 footprint were identified as possible downstream targets critical for the transition from cervical to thoracic identity.

2.10.2. CRISPR-ON to activate promoters

2.10.2.1. Primer design

Potential Cas9 target sites (18-20 nucleotides) were identified by scanning manually the promoter region of our gene of interest for PAM sequences (NGG). Identified sequences were then surveyed genome-wide using BLAT to ensure that the selected sequence has only a single

hit in the genome and then run through oligo-calc to identify any potential self-annealing sequences. Forward and reverse oligos for each guide recognition sequence (18-20nt) directly preceding the PAM (NGG), but not include the PAM itself, were designed. Sequences including a G base at the 5' end were preferentially selected for, as this is required for polIII to initiate transcription from a U6 promoter. In the case for sequences where this was not possible, a G residue was added upstream of the selected recognition sequence. Each of the sgRNA sequences used (Appendix 1) was also designed to include flanking BsmBI sites and corresponding overhangs for Golden Gate-based cloning into the U6 vector, where multiple fragments of DNA can be assembled by using combinations of overhang sequences in a single PCR step.

The oligo design is indicated in the figures below.

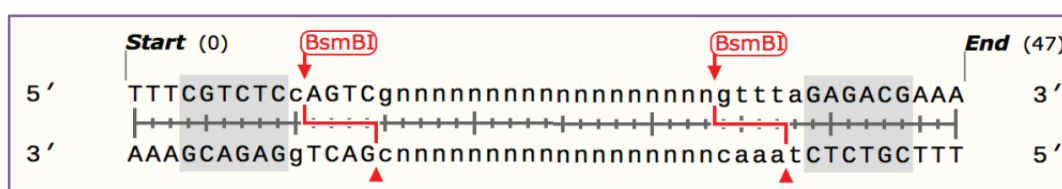


Figure 2.6: Guide RNA oligo design



Figure 2.7: U6.3 Mini Vector gRNA cloning site

2.10.2.2. Plasmid construct

As soon as the ordered primers arrived, they were resuspended in ST_{low}E at 100μM. 0.5μl of both the forward and the reverse primers were added into 49μl of ST_{low}E and then annealed at 94°C for 5 min. Following this, the product was cooled from 64°C to 34°C, with the temperature declining at 1°C/min for 30 min. The product was then run on a 2% agarose gel to confirm the correct size and then the concentration confirmed using a NanoDrop spectrophotometer. Annealed oligos were then cloned into the U6.3 vector using the following reagents:

Table 2.21: sgRNA plasmid reaction setup

Reagent	Volume (μl)
75ng pcU6.3 vector	x
4ng annealed oligos	x
10X T4 Ligase buffer	2
BsmBI (NEB; RO580)	1
T4 DNA ligase (NEB; MO202)	1
H ₂ O (SIGMA; W5402) to final 20μl volume	x

The modified Golden Gate reaction conditions were as follows:

Table 2.22: sgRNA plasmid PCR conditions

PCR step	Temperature (°C)	Time (min)	Cycles
Activation	37	5	10
Annealing	16	10	
Extension	50	10	1
Final extension	80	10	1

5μl of each sample were then directly transformed into 50μl DH5-alpha competent *Escherichia coli* cells (section 2.3.3.3) and the final resuspended cells were then plated on selective LB-agar plates (Tetracycline 10μg/μl) overnight at 37°C. Following another overnight culture in 5ml LB medium, plasmid mini-preparations were performed according to manufacturer's instructions (Qiagen; SBS27106) and verified with sequencing using the universal U6-FWD primer: GATCAAGCCTGATTGGGAGAAA.

2.10.2.3. Plasmid injection and electroporation

Plasmids (Appendix 1) were injected *ex ovo* into stage HH4 embryos using the method described above (section 2.9.3). Following observation under a fluorescence stereomicroscope, some embryos were immediately snap frozen for RNA extraction (section 2.1.2.3) for RTqPCR. Other embryos were fixed in 4% PFA and dehydrated in a series of methanol dilutions (section 2.1.3) for use in future WISH experimentation.

CHAPTER 3: RNA-SEQUENCING ACROSS THE C-T BOUNDARY

Studies of RNA and the transcriptome are of great importance in providing functional information and unravelling the genetic mechanisms that underlie evolution and development, as well as complex disorders and diseases. The advent of next generation sequencing has allowed the development of an assortment of technologies for interrogating aspects of the genome, one of which is high-throughput RNA sequencing (RNA-Seq). This technology allows rapid, relatively cheap, and accurate quantification of transcripts at a genome-wide scale. This study uses RNA-sequencing to explore the change in the transcriptome-wide profiles of the developing somites in the embryonic chick as they transition from cervical to thoracic axial identity.

3.1. RNA-Sequencing

RNA-Seq is a recently developed transcriptome profiling technology that utilizes next-generation sequencing platforms (Kadakkuzha *et al.*, 2016; Wang *et al.*, 2009). Here, a population of RNA is converted into a library of cDNA fragments with adaptors attached to one end or both ends. Each molecule is then sequenced in a high-throughput manner to obtain short sequences. Following sequencing, the resulting reads are either aligned to a reference genome or reference transcripts, or assembled de novo without the genomic sequence to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene (Wang *et al.*, 2009) (Fig 3.1).

There are some drawbacks to RNA-Seq, the most renowned of which include GC-biases and the issues associated with short read sequencing. GC-bias is a phenomenon observed in most NGS systems as a natural result of sequencing by synthesis, the result of which (in RNA-Seq) is an underrepresentation of both GC-rich and GC-poor regions (Risso *et al.*, 2011). One major disadvantage of short read sequencing includes the inability to cover the entire transcript isoform. This means that analyses of elements such as splicing events relies on junction-spanning reads and probabilistic models, which would not be as heavily relied upon if the entire RNA or fragment could be sequenced. Although these drawbacks are not insignificant, the many advantages of RNA-Seq have made it a widely used method. Additionally, there are several new RNA-Seq technologies either under development or newly released, which show promise in resolving many of these issues.

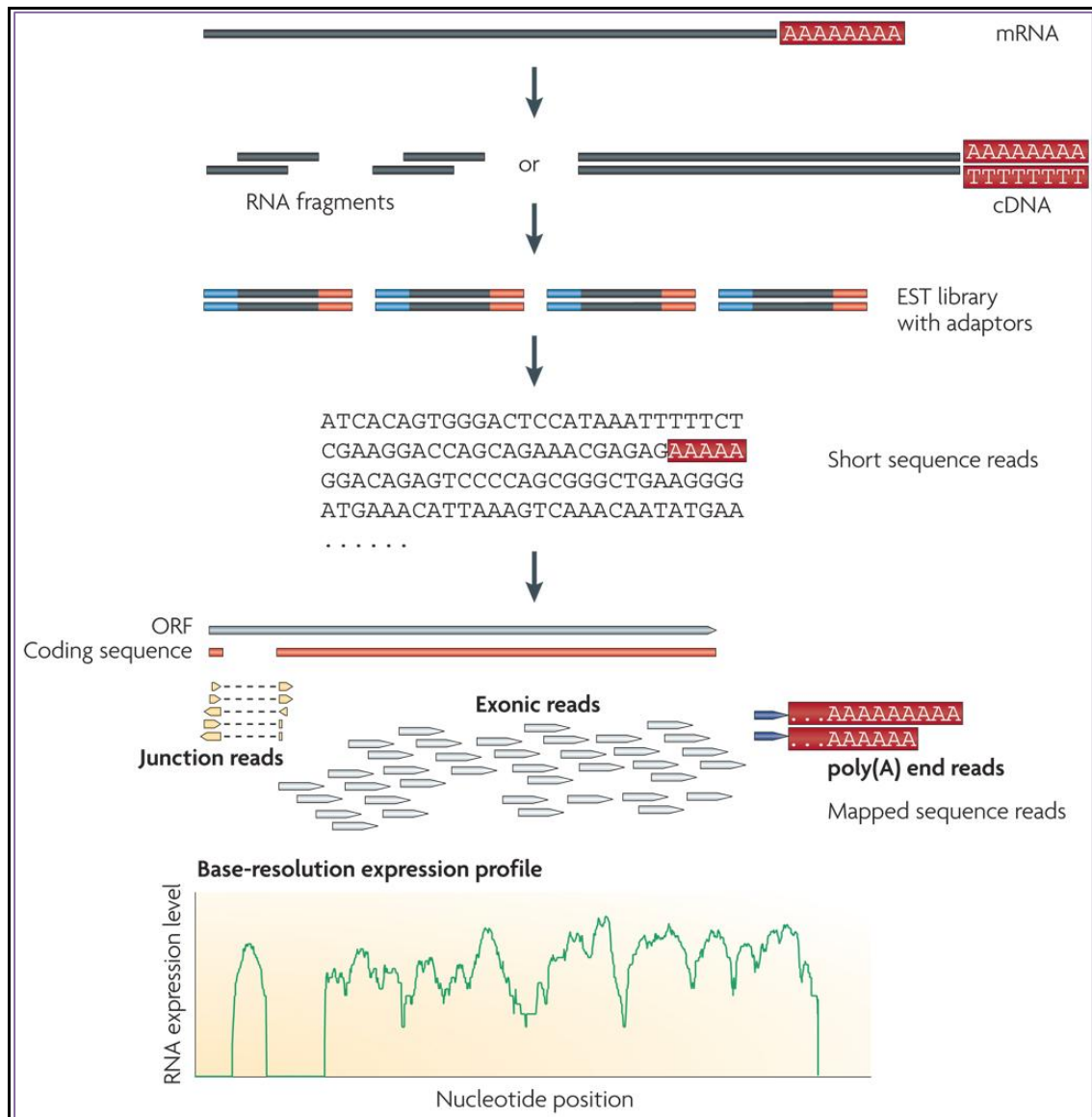


Figure 3.1: A typical RNAseq experiment. Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; yeast ORF with one intron is shown. Figure taken from (Wang *et al.*, 2009).

The raw data received from sequencing is commonly reported in the form of a FASTQ file, a text-based format containing strings of biological sequence and subsequent quality values corresponding to each sequence letter (Cock *et al.*, 2010). Each quality value conveys the probability that the corresponding base call is correct. This is based on the historical Phred score, a Sanger sequencing metric algorithmically determined from electropherogram peak resolution

and shape, the Illumina equivalent of which is largely the same, based on parameters relevant to sequencing chemistry (Illumina Inc. 2011). The quality values can be used for applying quality control procedures and ensuring that the data are sufficiently robust. Alignment of these reads to the complementary sequence of a reference genome (or de novo assembly without a reference) then allows for the counting of numbers of reads overlapping genomic features. The genome-wide digital quantification of expression is what provides RNA-Seq with its many advantages over the spot intensity quantification used by hybridization-based methods.

RNA-Sequencing allows full genome-wide quantification analysis, not limited to those features chosen to be included in a microarray, and not limited to quantification ceiling by probe number limits. RNA-Seq also allows for quantification of novel transcripts, whose sequences are unlikely to be included in microarrays, such as noncoding RNAs, and for which (although lncRNA and miRNA microarrays exist) detecting low expression levels can exhibit poor sensitivity in microarrays. Finally, RNA-Seq allows for the more accurate estimation of splicing variants, with reads spanning multiple exons. There is a microarray equivalent for investigating splicing, which is the exon junction microarray. This uses probes specific to potential splice sites. However, neither Illumina RNA-Seq nor exon junction microarrays provide full length transcript information, and so all isoform quantifications are based on probabilistic models. The advantages of Illumina RNA-Seq are its abilities to interrogate exons.

Overall, RNA-Seq is a powerful tool for interrogating transcriptomic signatures and this can successively provide a much greater understanding of the underlying mechanisms required for anterior-posterior patterning as well as for unveiling the evolutionary mechanisms behind this. As mentioned previously, there are also many advantages provided by RNA-Seq compared to other technologies. These advantages can be useful in identifying novel transcripts, can also be used for the building of more accurate gene regulatory networks.

3.2. Transcriptional Profiling of Cervical and Thoracic Somites

The six most recently formed individual somites, equivalent to somite numbers 17-22, were collected in quadruplicate from Hamburger-Hamilton stage 14 (HH14) chick embryos in order to capture in total the spatiotemporal transition from cervical to thoracic identity, which occurs across somites number 19 and 20 in the chick. At this stage, the three most posterior thoracic somites are epithelial, but the more anterior cervical somites, as described in Chapter 1, are differentiating and becoming compartmentalised, with a primary myotome beneath the dermomyotome and a sclerotome ventrally. This change in maturation of the somites would skew our data, resulting in an increased number of genes calling as differential, which are

involved in maturation of the somite rather than in axial identity. To counteract this, another dataset was added to this, which included the addition of the three most recently formed somites from a stage HH13 chick embryo (somites number 17-19). These epithelial somites were added for comparison in order to remove any genes called as differential that are not involved in the fixation of axial identity (Fig. 3.3).

3.2.1. Quality Control Procedures

The Quality Control (QC) procedures executed throughout the RNA-Seq analysis serve the purpose of assessing the quality of the data, as well as correcting the raw sequence reads for any problems or biases observed. Although most sequencers generate QC reports themselves, these are usually focused on problems from the sequencing itself. Further initial assessment of the data was carried out using FastQC (Andrews, 2010), which is a programme that provides a modular set of analyses to give an impression of the overall quality of the raw sequence reads through multiple visualizations. The modules returned by FastQC provide information relevant to the sequencing itself, but also relating to problems that may occur in generating the starting library material.

In incidences of poor base quality (phred < 20), the FASTQ Quality Trimmer was used to cut reads down if their quality scores dropped below a certain threshold towards the end of the read. The FASTQ Quality Filter was also used to filter reads of low quality throughout. In incidences of adapter sequence contamination, Cutadapt (Martin, 2011) was used for cutting reads down based on the presentation of adapter sequences.

3.2.2. Read Alignment and Quantification

Following quality control procedures, the raw reads were aligned to the GalGal6 chicken genome assembly (Ensembl Release 99) (Yates *et al.*, 2020) with the splice-aware short read aligner GSNAP (Wu & Watanabe, 2005). GSNAP is a seed-and-extend short-read aligner that can be used for both single- and paired-end reads as short as 14nt. As of 2012, there are at least 60 aligners available for use in high-throughput sequencing alignment (Fonseca *et al.*, 2012), and so selection of a suitable aligner can be difficult. For RNA-Seq it is important to use an aligner which is splice-aware and allows for long distance alignment of reads across splice junctions. Splice-aware aligners accommodate junction-spanning reads through splitting the reads into smaller segments.

This is executed through one of two methods depending on the aligner, either “exon-first” alignment or “seed-and-extend” alignment. Exon-first alignment begins by aligning the entire read and then searching for spliced alignment for the remaining reads, whereas seed-and-

extend alignment uses short sub-sequences, or seeds, determined by breaking down the sequences of the reads and aligning these to the genome. The seed-and-extend aligners are slower but offer the advantage of higher sensitivity (Benjamin *et al.*, 2014). Although many popular aligners use exon-first alignment (For example TopHat2; Kim *et al.*, 2013), a seed-and-extend aligner is used throughout this thesis due to the increased sensitivity associated with these. Namely, the Genomic Mapping and Alignment Program (GMAP) (Wu & Watanabe, 2005) and its Genomic Short-read Nucleotide Alignment Program (GSNAP) are used.

Following read alignment, the data can be quantified through counting the number of reads that overlap features of the genome. For this, HTSeq (Anders *et al.*, 2015) was used for counting reads either contained within gene boundaries, or within specific transcript boundaries when implementing splicing analyses and determining differential exons/isoforms. Gene/transcript boundary information was acquired in the form of Gene Transfer Format (GTF) files from the Ensembl genome database (Yates *et al.*, 2020), which provide gene structure information conveying to the relevant reference sequence.

3.2.3. Early Analyses

Differential expression analysis is that which seeks to identify genes that are expressed at different levels in different conditions. This testing was carried out using the standardized pipeline developed in the Hearty lab at the Earlham institute, with the process automated to perform the analysis across all paired combinations of conditions. Differential expression analysis was performed with the DESeq2 package (version 1.36) (Love *et al.*, 2014) within the RStudio programming environment (version 1.3.959). An adjusted p-value of ≤ 0.05 was used as the significance threshold for the identification of differentially expressed genes. This particular package is implemented here as it was found to be the most robust and reproducible package available. DESeq2 aims to test whether the measurements for the expression of a given gene in one condition comes from a significantly different distribution to another. The null hypothesis of this is that there is no difference, and that in the given gene the expression values in one condition come from the same distribution as another condition. This involves some analysis prior to the statistical test, which approximates a suitable distribution of the expression data. As the expression data is based on count data, a Poisson distribution would normally be most suitable. However, RNA-Seq datasets suffer from an overdispersion problem, where there is the presence of a greater than expected variance at larger mean counts, and so are more suited to the negative binomial distribution. DESeq2 uses this model in addressing the underlying hypothesis with a customised statistical test that is a variant of the Fisher's exact test but more suited to this type of distribution.

Normalization of the data is key in ensuring differences between samples are due to true biological variation as opposed to a simple difference in library size. Classically, this could be done through dividing all values by the total read count, thus giving resultant values totalling the same in each sample. However, DESeq provides a more complex and better fitting solution (Dillies *et al.*, 2013). It normalizes the count data with a customised method aimed at maintaining the assumption that most genes are not differentially expressed. DESeq calculates a scaling factor for a given sample as the ratio between the median read count value across all genes within that sample and the geometric mean across all samples. This is based on the idea that genes which are not differentially expressed should have similar read counts across all samples, and a resultant ratio of one. When estimating the dispersion for the negative binomial distribution to be used in statistical testing, DESeq uses the maximum individual dispersions and the dispersion-mean trend (Anders *et al.*, 2013), resulting in a more conservative result than from alternative differential expression packages, such as EdgeR (Robinson *et al.*, 2010). Due to the large number of genes that were identified as differential, a more conservative approach is preferable to help narrow down the search for key genes involved in securing somite identity across the cervical-thoracic boundary.

3.2.4. Gene Ontology

A common method used alongside RNA-Seq analyses is enrichment analyses, such as Gene Ontology (GO) enrichment. There are many functional enrichment analysis tools available and widely used for these analyses (Huang *et al.*, 2009). These rely on various functional annotation databases (such as ontologies or pathways) and their associations with each gene across the genome. A given list of genes can be profiled with these methods by testing for enrichments of functional annotations. Gene set enrichment analysis for differentially expressed genes in this dataset was performed using the g:GOST programme within the g:Profiler web browser (Raudvere *et al.*, 2019). This analysis was applied to the differentially expressed genes identified.

3.2.5. Hierarchical Clustering

Clustering is a statistical classification technique in which the main objective is to form groups or clusters from a set of objects, and for each cluster to contain objects showing a greater similarity with one another than with objects in other clusters (Altman & Krzywinski, 2017). Clustering is an approach to data mining which is utilized in many fields, one of which is for elucidating patterns in large biological data sets such as those produced from next-generation sequencing. For example, in RNA-Seq clustering is frequently used to extract commonly occurring gene expression patterns observed over time-series data, where each cluster contains

a group of genes with similar expression patterns (Baumgart *et al.*, 2014; Cho *et al.*, 2014; Klepikova *et al.*, 2015; Marthandan *et al.*, 2015; Oberstaller *et al.*, 2013). The clustering algorithm suggested for RNA-Seq is model-based clustering with a negative-binomial model.

3.3. Results and Discussion

RNA-Sequencing produced a mean of ~30 million reads per sample and there was no evidence of lane bias in the multiplex sequencing. FastQC returned no evidence of low-quality reads, and so read trimming and filtering were not deemed necessary. RNA-Seq analysis proceeded using various packages within the RStudio programming environment.

3.3.1 PCA

Principal Component Analysis (PCA) was applied to the expression count data for all genes across all samples (Table 3.1; 3.2) for visualization of sample relationships to one another (see Fig 3.3 for a cartoon representation). This was implemented with all count data, with stabilized variance through DESeq2.

Table 3.1: Samples for Differential 1 (D.1) of RNA-Seq.

Replicate	Sample Designation	Axial domain	Somite no.	Somite Age
1	S1	Cervical	n17	Maturing
1	S2	Cervical	n18	Maturing
1	S3	Cervical	n19	Maturing
1	S4	Thoracic	n20	Epithelial
1	S5	Thoracic	n21	Epithelial
1	S6	Thoracic	n22	Epithelial
2	S7	Cervical	n17	Maturing
2	S8	Cervical	n18	Maturing
2	S9	Cervical	n19	Maturing
2	S10	Thoracic	n20	Epithelial
2	S11	Thoracic	n21	Epithelial
2	S12	Thoracic	n22	Epithelial
3	S13	Cervical	n17	Maturing
3	S14	Cervical	n18	Maturing
3	S15	Cervical	n19	Maturing
3	S16	Thoracic	n20	Epithelial
3	S17	Thoracic	n21	Epithelial

3	S18	Thoracic	n22	Epithelial
4	S19	Cervical	n17	Maturing
4	S20	Cervical	n18	Maturing
4	S21	Cervical	n19	Maturing
4	S22	Thoracic	n20	Epithelial
4	S23	Thoracic	n21	Epithelial
4	S24	Thoracic	n22	Epithelial

Table 3.2: Samples for Differential 2 (D.2) of RNA-Seq.

Replicate	Sample Designation	Axial domain	Somite no.	Somite Age
1	S25	Cervical	n17	Epithelial
1	S26	Cervical	n18	Epithelial
1	S27	Cervical	n19	Epithelial
2	S28	Cervical	n17	Epithelial
2	S29	Cervical	n18	Epithelial
2	S30	Cervical	n19	Epithelial
3	S31	Cervical	n17	Epithelial
3	S32	Cervical	n18	Epithelial
3	S33	Cervical	n19	Epithelial
4	S34	Cervical	n17	Epithelial
4	S35	Cervical	n18	Epithelial
4	S36	Cervical	n19	Epithelial

In the initial analysis, a significant batch effect could be seen occurring in both differential conditions tested (Fig 2.4; Fig. 3.2). I believe this batch effect was caused by the timing of the experiment. Although all somites were removed on the same day, about an hour and a half was needed to complete the isolation procedure for the first two embryos (replicates 1 and 2) before proceeding with the next two (replicates 3 and 4), which is where the divide is occurring on the PCA plot. This was further confirmed when comparing count number for typical genes known for their involvement in the segmentation clock of the chick, such as HAIRY1/2 and LFNG, which showed variation between replicates 1/2 and 3/4. This suggests that the development of the final two embryos was further along than those of the first two embryos dissected. Due to the differing development, many genes required for the timing of somitogenesis were called as differential, resulting in the batch effect. To remove this, I regressed out the batch effect by

using `Combat_seq` from the BioConductor package (version 1.30.10), which provides adjusted data by mapping the original data to an expected distribution if there were no batch effects.

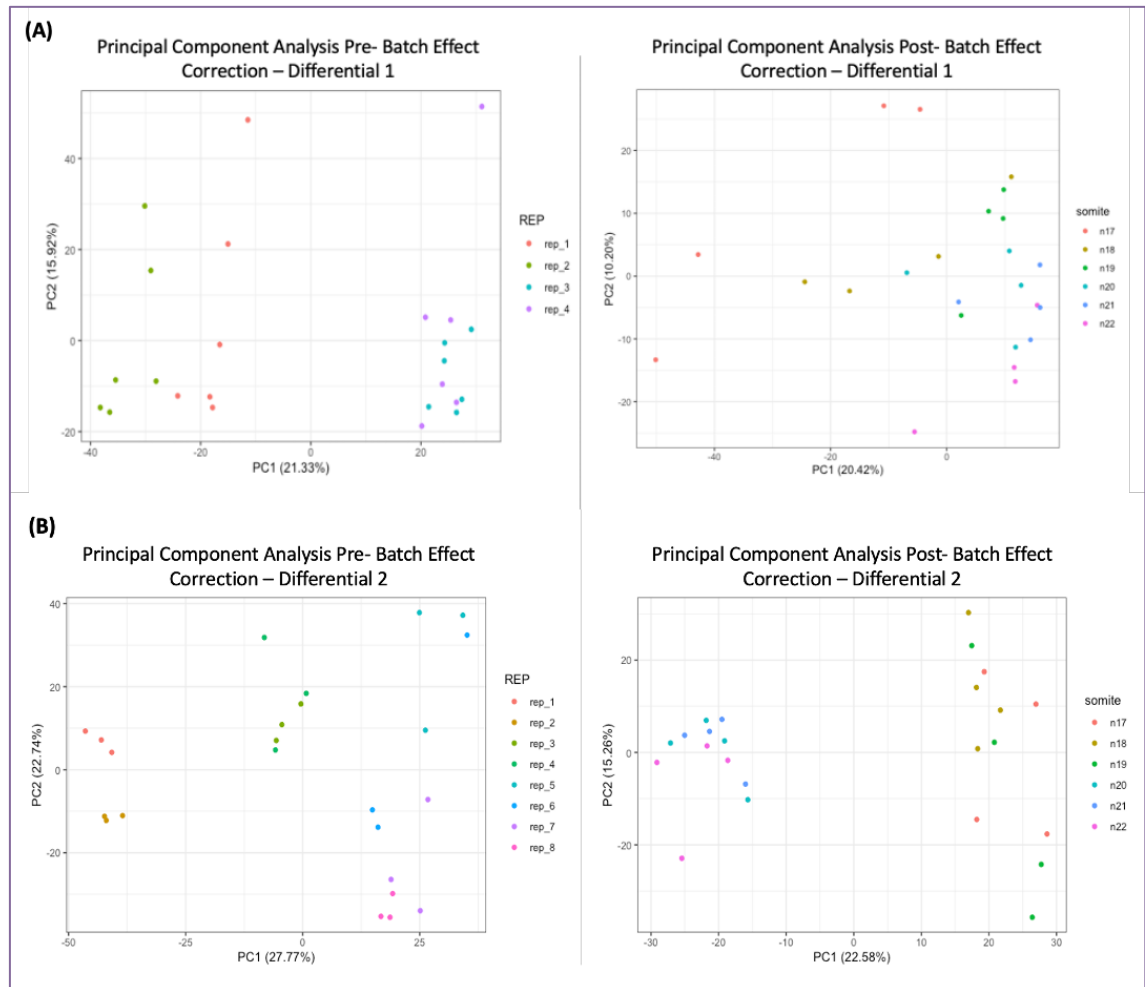


Figure 3.2: Principal Component Analysis (PCA) Plots. Visualisations of the PCA plots both before and after regression for Differential 1 (A) and Differential 2 (B). Pre- Batch Effect points are coloured by replicate number. For differential 2, replicates 1, 3, 5 and 7 refer to the thoracic somites from stage HH13+ embryos and replicates 2, 4, 6 and 8 refer to the cervical somites from younger stage HH12 embryos (see Fig 3.3 for a cartoon representation). Post- Batch Effect points are coloured by somite number.

Following regression of the batch effect, the PCA plot for Differential 1 showed that the cervical and thoracic samples cluster into two groups, with the somites developing on the cervical-thoracic border (somites no. 19 and 20) clustering together, indicating they have the most similar molecular makeup to each other, clustering together. As the somites move further away from that border, they begin to cluster into more distinct groups, with the most anterior cervical somites number 17 and the most posterior thoracic somites number 22 having the widest difference 20.42% in PC1. This is due to biological variance. If the PCA plots were generated comparing only the somites furthest away from each other (somites 17 vs. 22, or somites 17/18 vs. 20/22), two much more distinct groups would appear on the plot (Appendix 3). Instead, due

to the genetic similarity of the two most centred somites above the C-T border (somites 19 and 20), we can see a gradual change in genetic makeup across the samples.

An even more distinct clustering of the two groups can be seen within Differential 2 when comparing epithelial somites only with a difference of 22.58% in PC1. This greater separation of the two groups, cervical and thoracic, is also due to both groups originating from different embryos, unlike the first condition, where the six somites were extracted from the same embryo (Table 3.1).

As the somites mature, we observe a separation between replicates in PC2 for both Differential 1 and Differential 2 (10.2% and 15.26% respectively). This can be explained by the subtle differences in the somitogenesis clock and wavefront cycle of the developing somites, as explained in Chapter 1, that somewhat still remain even after regression of the batch effect as the difference in developmental stage of the individual somites made an even regression across samples impossible. Due to the difference in expression level of some typical genes known for involvement in the Clock and Wavefront model seen in the RNA-Seq dataset, I believe that residual expression of these genes causes this change.

3.3.2 Differential expression

Differential gene expression analysis comparing cervical and thoracic somites extracted from the same stage HH13+ embryo across the C-T boundary (Differential 1; Fig. 3.3A) revealed 1105 differential genes, with 466 genes up-regulated within thoracic somites and 639 genes down-regulated within thoracic somites.

When comparing epithelial, cervical-level somites from a stage HH13 embryo to epithelial, thoracic-level somites from a HH14 embryo (Differential 2; Fig. 3.3A), a total of 6082 genes were identified as differential, with 3299 genes up-regulated in thoracic somites and 2783 genes down-regulated.

By adding a third comparison, where differential genes were identified between the epithelial cervical somites from Differential 2 and the maturing cervical somites from Differential 1 (Differential 3; Fig. 3.3A), I could identify all differential genes that are involved in somite maturation only and remove those from consideration. After removal of this comparison, I overlapped Differential 1 and Differential 2 gene lists, thus identifying a much narrower list of differential genes that may be responsible for the change in axial identity from cervical to thoracic fate (Fig. 3.3B). The total number of differential genes identified was reduced to 343,

with 153 common elements up-regulated in thoracic somites and 190 common elements down-regulated in thoracic somites.

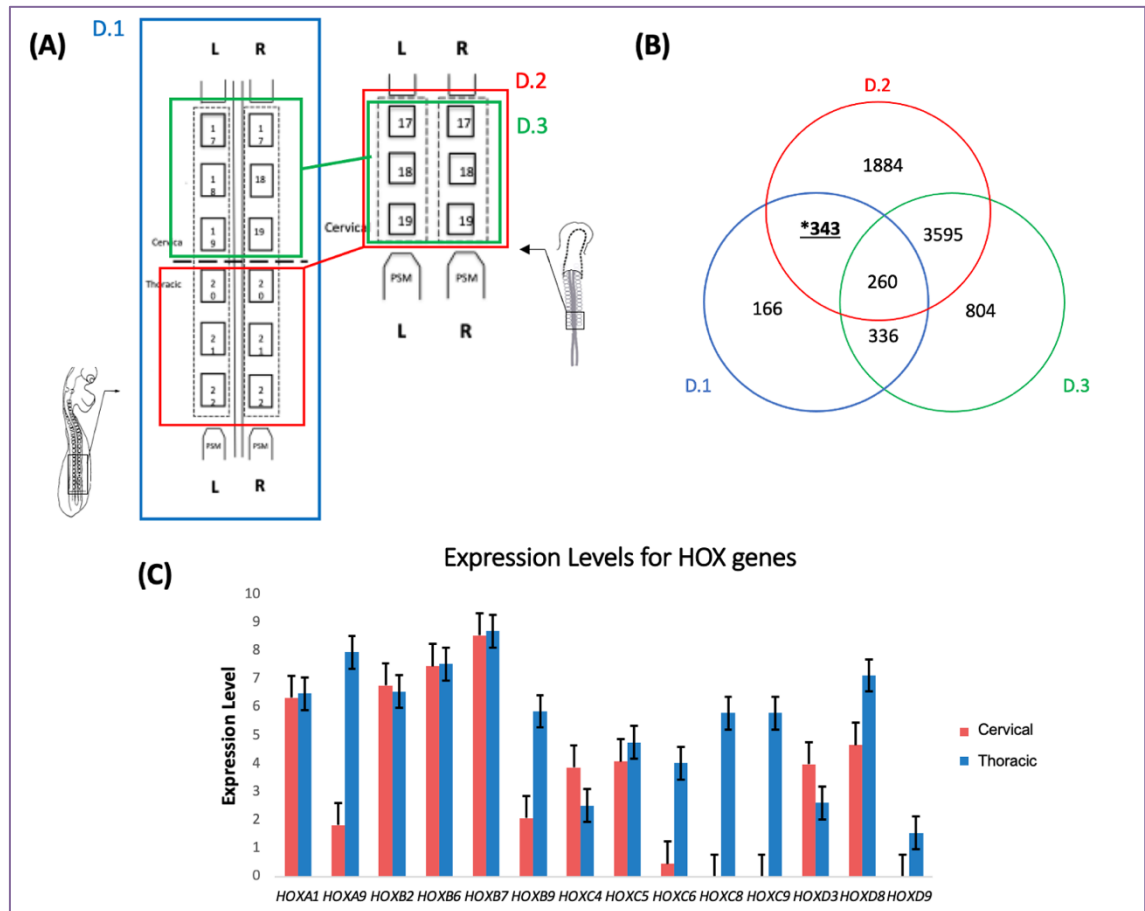


Figure 3.3: Transcriptional Profiling of Somites across the C-T Boundary. (A) Schematic representation of the experimental setup for three differential conditions comparing somites of differing developmental stage across the C-T boundary (differential 1 (D.1) marked in blue), somites of the same epithelial stage from different staged embryos across the C-T boundary (differential 2 (D.2) marked in red), and somites of different developmental stage but of the same cervical axial level (differential 3 (D.3) marked in green). (B) Venn diagram of all differentially expressed genes across the three comparisons. The starred intersection between D.1 and D.2 represents all 343 overlapping genes predicted to be responsible for positional information only. (C) Differential RNA expression of Hox genes responsible for axial patterning in the cervical somite samples from Differential 1 (red) – thoracic (blue) boundary.

From this DEseq analysis, the topmost differentially expressed genes across the two domains identified were the HOX genes. Specifically, *HOXC8* and *HOXC6* arose as the most differential. Other Hox genes, such as *HOXA9*, *HOXB9*, *HOXD3*, and *HOXD9*, also appeared very high in the list of differentially expressed (DE) genes, as shown in the bar plot above, comparing their levels of expression in the cervical domain vs the thoracic domain (Fig. 3.3C). The only genes that appeared with no expression in cervical somites versus a sudden increase in expression in

thoracic somites were the more posterior expressing Hox genes. All other genes identified above the threshold for cut-off as differential were consistently expressed within both axial levels and showcased a slight log-fold change in expression, rather than a sudden onset or offset of expression.

3.3.3 Heat maps

Heatmap visualization and hierarchical clustering were also performed to corroborate the relationships seen in the PCA findings. Visualizations were implemented with all normalized count data (Fig. 3.4A), with stabilized variance through DESeq. Hierarchical clustering was performed with the complete (maximum) linkage method and using Euclidean distance measures.

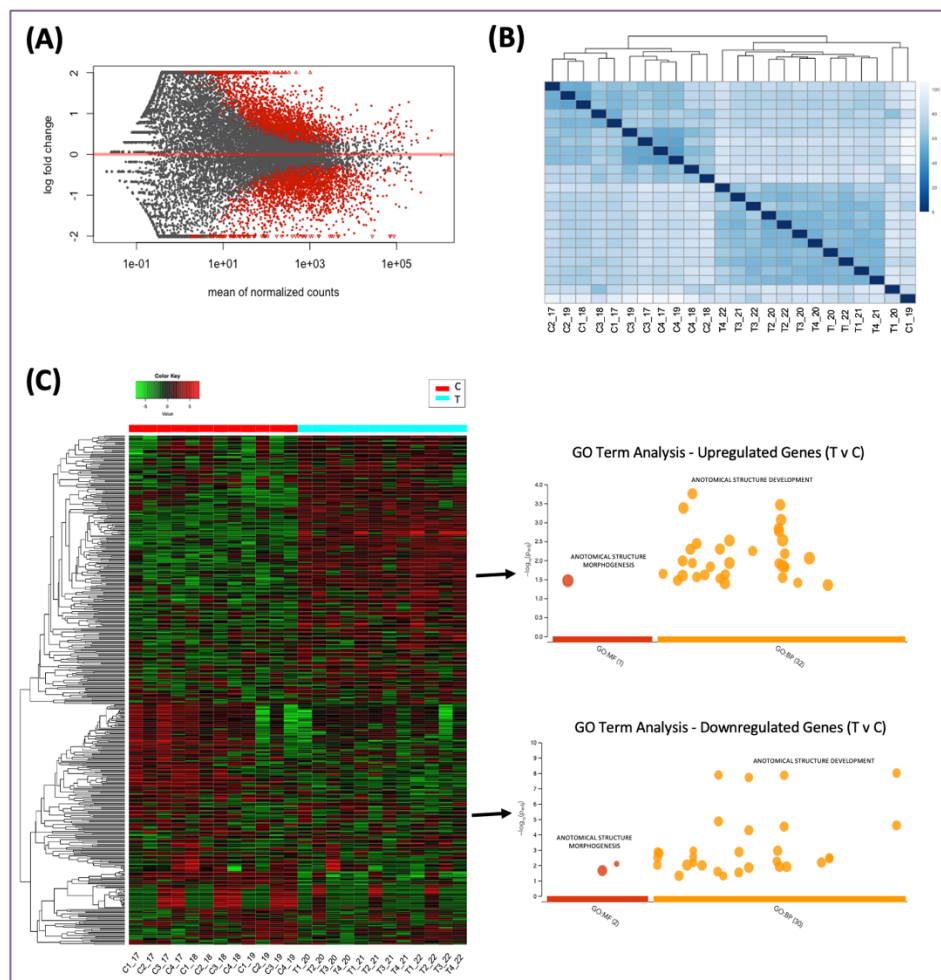


Figure 3.4: RNA-Seq Data Quality Check and Visualisation. (A) MA plot providing a global view of the differential genes. (B) A heatmap with conjoining hierarchical clustering dendrogram for all of the count data of all of the replicates of the Differential 1 data set. (C) Heatmap depicting k-means linear enrichment clustering for all 1105 differential genes across the C-T boundary in Differential 1 with corresponding dot plot visualisation of GO Term Analysis for genes upregulated in thoracic somites (top plot) and genes downregulated in thoracic somites (bottom plot).

The relationships seen in the PC analysis are further confirmed by visualization through heatmaps and by hierarchical clustering of the data, with the two major clusters formed depicting separation of the cervical and thoracic somites (Fig. 3.4B). With the exception of some similarities between somite samples 19 and 20 (Fig. 3.4B; C), the heatmaps show an expected separation of replicates.

3.3.4 Gene Ontology

The functional clustering by gene ontology (GO) terms of differentially expressed genes across both stages reveals enrichment of biological processes involved in anatomical structure morphogenesis and development for both upregulated and downregulated genes (Fig. 3.4C). Genes involved in anterior-posterior pattern formation, bone morphogenesis, and muscle differentiation were also identified. Classic markers for chondrogenesis and cartilage condensation identified include the TFs, SOX9, SOX5, SOX6, BMP7 and PAX1. These are expressed in an opposing fashion along the axis include the FGF and Wnt signalling pathways. Consistent with the known patterns of expression, Wnt and FGF pathway components are highly expressed in thoracic samples and the Notch signalling pathway, which is more highly expressed in cervical samples.

3.3.4 Additional Analyses

Hierarchical clustering analysis was carried out to further partition, or cluster, differential genes together based on similar trend of expression pattern across the 6 somites spanning the cervical-thoracic boundary. The first step in hierarchical clustering is to create a dendrogram (Fig. 3.5). This groups the differential genes based on their trends of expression across the somites. For example, the largest separation, the top two groups where the red line is, is separated by genes that increase in expression from cervical to thoracic somites, versus those that decrease in expression across the boundary. As you go lower in the dendrogram, it creates more and more groups, clustering genes into small very specific trends of expression.

For greater comparison, and to reduce complexity caused by the somites overlapping the boundary, somites numbered 19 and 20 were removed from this analysis, comparing only somites numbered 17 and 18 vs 21 and 22. By merging clusters of genes together based on eigengene similarity, a total of 5 clusters were identified amongst the differential genes, which were plotted in a line graph for greater visualisation.

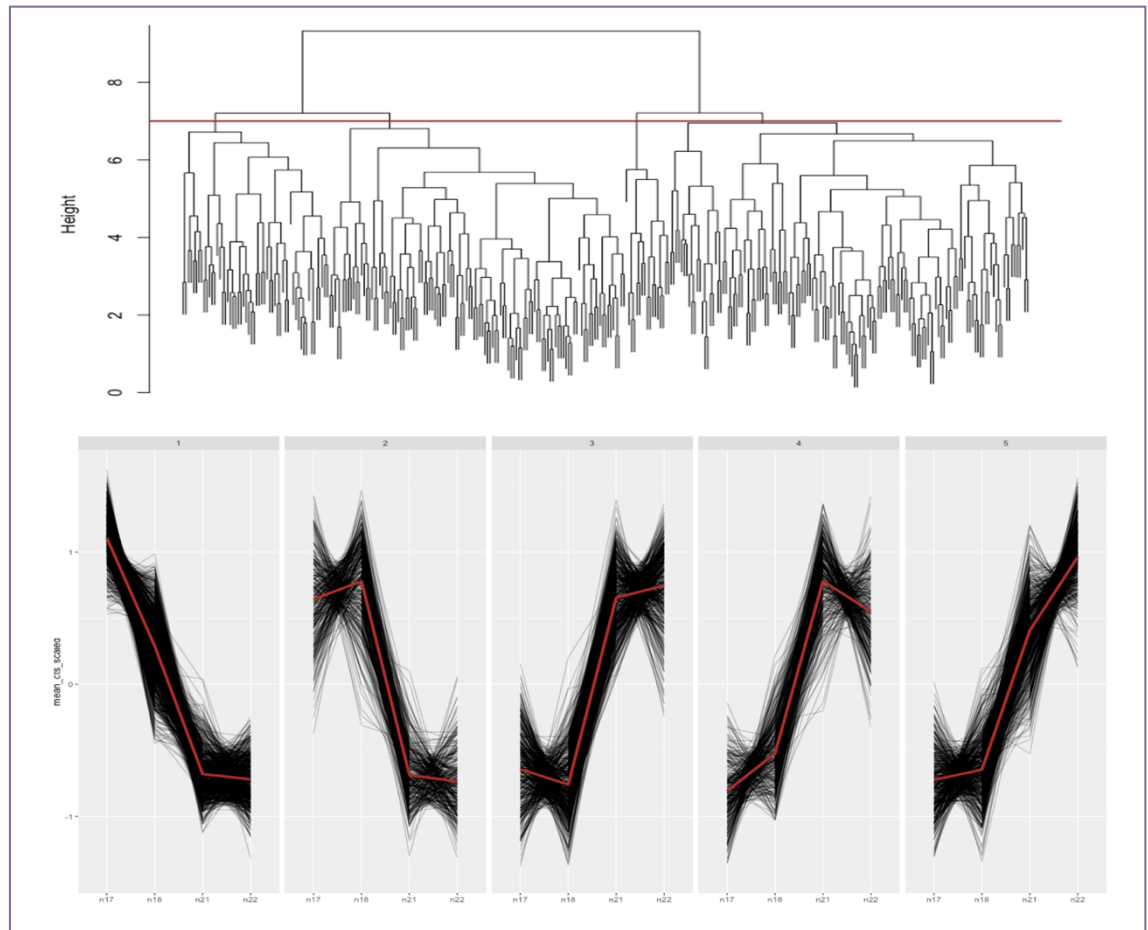


Figure 3.5: Hierarchical clustering of RNA-Seq data. Dendrogram of all differential genes from the RNA-Seq dataset generated from Differential 1 clustered based on eigengene similarity with corresponding line graphs displaying standardized gene expression across each of the 5 clusters identified by the red line across the dendrogram above. Each plot represents a single cluster, and each line represents the expression of a single gene over the course of developmental timing. Expression data was standardized to a mean of zero, and a standard deviation of one.

Of these clusters, number 3 and 5 are associated with similar expression patterns to *HOXC6* and *HOXC8/9* respectively, both of which are positively associated with developmental time. Many of the genes within both clusters are associated with anterior-posterior patterning and anatomical structure development. For example, *HAPLN1* is a gene known for its structural support in cartilage formation and *PDGFA* is a growth factor expressed in skeletal cells, or osteoblasts.

By clustering the dataset, the large number of differential genes of focus is greatly reduced and I can then preferentially select for genes that follow the same trend in expression as other genes known to be important in determining this axial identity, such as *HOXC6*. By combining this information with *in silico* footprint analyses from the ATAC-Seq dataset, which will be discussed

in Chapter 4, I aim to identify a downstream network of genes vital for implementing cervical/thoracic identity following Hox activation.

3.4. Conclusion

This study pursued the identification of the underlying molecular signatures underlying change in axial identity from cervical to thoracic, reporting genome-wide expression profiles through RNA-Seq in the developing chick embryo. PCA revealed a gradual change across the C-T boundary, resulting in distinct clusters between the most anterior cervical somite and the most posterior thoracic somite and illustrated the segregation of the two axial identities.

I report a library of differentially expressed genes, narrowing the search to 343 marker genes which, following stringent selection criteria and depending on further experimental testing, may provide important insight and information on the mechanisms underlying the transition from cervical to thoracic identity in vertebrate axis development. Many of these markers are already known to be involved in mechanisms of development relating to anterior-posterior patterning and anatomical and/or bone morphogenesis through various gain/loss of function experiments. Future work is necessary to determine the necessity of these genes for cervical/thoracic identity and to confirm and build the gene regulatory network surrounding this phenomenon with confirmation of interactions between these genes and the Hox genes. These markers provide a resource for tracking cell fate acquisition and guide us towards further investigations into the phenotype of the systems. For example, Hox knockdowns/gain of function studies followed by RTqPCR to track a change in expression for these markers could be extremely revealing and may uncover previously unknown genetic interactions.

Overall, this study provides initial insights into the molecular signatures underlying the transition from cervical to thoracic somite identity. These markers can form the basis of future studies, important in investigations into axial development, as well as in investigating changes in genetic mechanisms, perhaps contributing to further study into the genetic differences underlying evolution of various vertebrate species.

CHAPTER 4: IDENTIFICATION AND VALIDATION OF HOX *CIS*-REGULATORY ELEMENTS ACROSS THE C-T BOUNDARY

4.1. Introduction

Modifications in gene regulation are driving forces in the evolution of organisms. Part of these changes involve *cis*-regulatory elements (CREs), which contact their target genes through open chromatin structures. Mechanisms of somite formation and patterning are largely conserved along the length of the body axis, but segments acquire different morphologies in part through the action of Hox transcription factors (Chen *et al.*, 2013). Identification and characterization of the molecular signatures responsible for the functional specificity of Hox genes and other correlated genes identified will help in our understanding of the patterning activity of these proteins (Casaca *et al.*, 2016).

The proper expression of Hox genes is necessary for the accurate patterning of the body plan. Understanding the interactions between *cis*-regulatory elements and Hox genes and how these elements subsequently trigger Hox functional activity requires the identification and characterisation of the factors that interact with them. By determining the developmental genetic basis of transcriptional regulation of Hox genes by the study of their *cis*-regulatory elements, crucial information regarding the establishment of axial specification may be uncovered.

The Assay for Transposase Accessible Chromatin with high-throughput Sequencing (or ATAC-Seq) is used to identify open chromatin regions along the eukaryotic genome where putative *cis*-regulatory elements (CREs), such as enhancers, may be acting to control transcription of neighbouring genes. Enhancer sequences are regulatory DNA sequences that, when bound by specific transcription factors, enhance the transcription of an associated gene.

Using ATAC-Seq, it is also possible to infer the positions of DNA-binding proteins within regulatory elements using *in silico* 'footprinting' technique, allowing identification of candidate transcription factors that are inducing activity of the enhancers. Regulation of transcription is the most common form of gene control, and the activity of transcription factors allows genes to be specifically regulated during development and in different types of cells.

Regulation of transcription is the most common form of gene control, and the activity of transcription factors allows genes to be specifically regulated during development and in different types of cells. The eukaryotic genome (DNA, RNA and protein) is packaged into structures called chromatin. These chromatin fibres are not condensed but can exist in either a compact or 'closed' form (heterochromatin), or a more accessible or 'open' form (euchromatin). Processes such as DNA replication, transcription, and recombination occur in the euchromatin, as these open chromatin regions allow biologically active regions – such as enhancers, promoters and other regulatory elements – to be accessible to different proteins, including for example the transcription machinery. Therefore, the nature of this packaging plays a central role in gene regulation, or epigenetics.

Bioinformatic analysis can also infer the positions of DNA-binding proteins using an *in silico* 'footprinting' technique. This footprinting technique allows scientists to identify potential DNA regulatory elements, such as enhancers, silencers, promoters, etc. which control transcription of neighbouring genes and are embedded upstream or downstream of a particular gene of interest in the genome. Enhancer sequences are regulatory DNA sequences that, when bound by specific proteins called transcription factors, enhance the transcription of an associated gene.

The ATAC-Seq procedure involves the use of a hyperactive transposase, transposase Tn5 on a sample of genomic DNA (Fig. 4.1). This cuts exposed DNA and simultaneously ligates adapters to the DNA fragments. Adapter-ligated DNA fragments are isolated, amplified by PCR, and used for sequencing. Open regions of the genome where DNA is more accessible will contain significantly more sequencing reads and thus form distinct peaks in the ATAC-Seq signal.

In this chapter, gene regulation of the Hox genes responsible for the transition from cervical to thoracic patterning along the anterior-posterior axis of the developing chick embryo (*HOXC5*, *HOXC6*, and *HOXC8*) is examined using ATAC-Seq. The aim was to identify putative enhancer elements and further examine the interactions required for their functional activity. Using *in silico* footprints and information from RNA-Seq data, downstream targets of *HOXC6* are also identified which may help uncover more of the mechanisms underlying development of thoracic vertebral features, such as ribs.

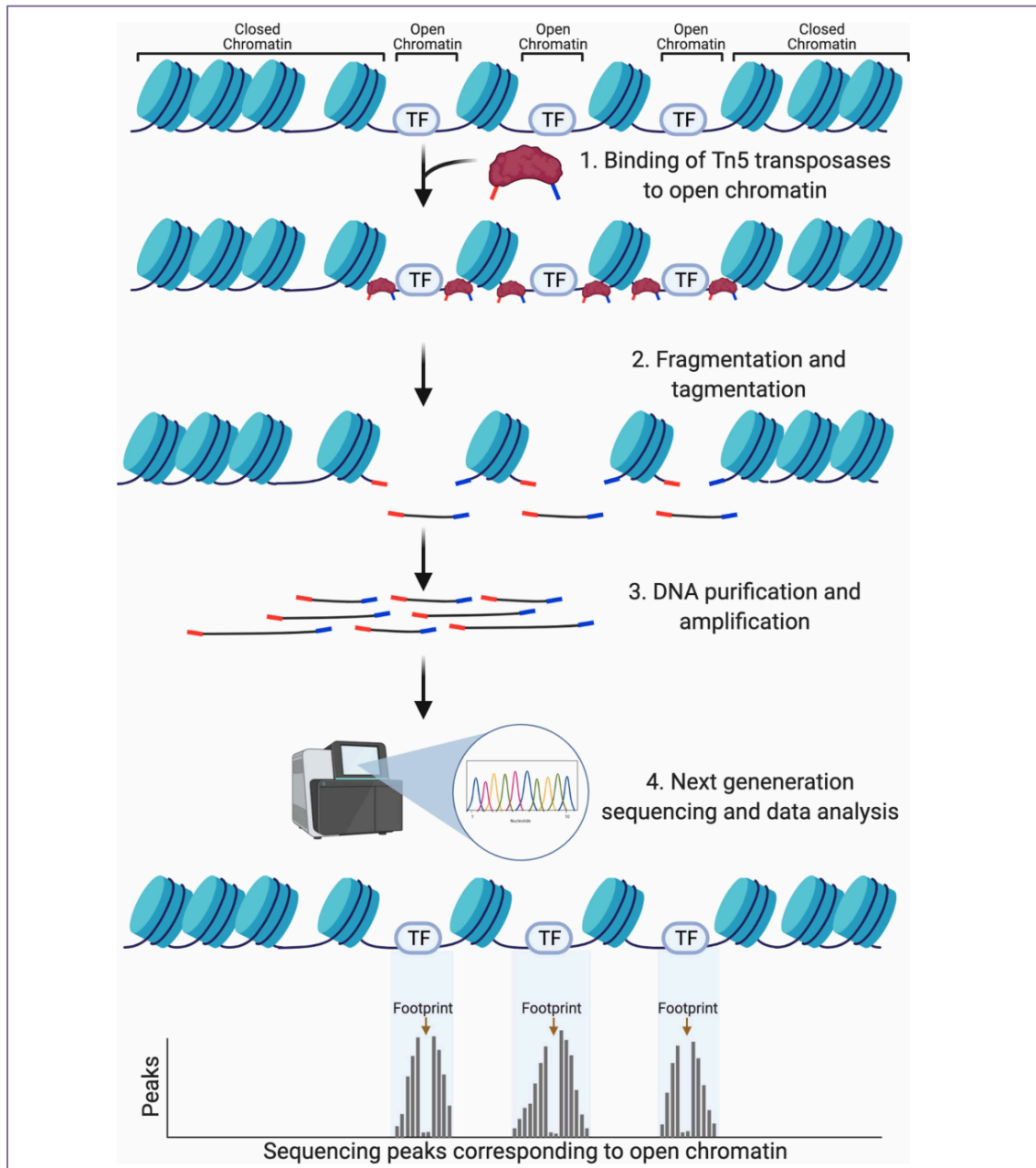


Figure 4.1: ATAC-seq methodology. (1) Tn5 transposase (brown) inserts sequencing adaptors (red and blue) into regions of open chromatin. Nucleosomes are shown in light blue, DNA-bound transcription factors (TF) are shown in grey. (2) The chromatin is fragmented and simultaneously tagmented with sequencing adaptors. (3) DNA is purified and PCR-amplified into a library using barcoded primers. (4) The library can then be analysed by qPCR or next-generation sequencing (NGS). Data analysis is performed and accessible regions of chromatin show as peaks. Within peaks, lower read coverage indicates TF footprints and allows prediction of TF binding *in silico*. Figure taken from (Smith *et al.*, 2022).

4.2. Results and Discussion

4.2.1. ATAC-Seq library preparation

To identify putative *cis*-regulatory elements that are responsible for the epigenetic control of HOX genes affecting the transition from cervical fated cells to thoracic fated cells, ATAC-Seq libraries were generated from the 6 most recently formed individual somites across the cervical-thoracic axial boundary. This mapped chromatin accessibility across the C-T boundary and was carried out in parallel with the RNA-Seq discussed in the previous chapter. To account for biological variability between samples and to allow for downstream statistical analyses, four biological replicate ATAC-Seq libraries were generated and sequenced at 30 million reads (150 pe) each. Distinct chromatin accessibility profiles were evident in the different somites, indicative of the dynamic progression of axial development. PCA showed a high reproducibility between biological quadruplicates of each sample type.

To assess in an unbiased fashion the way in which cervical and thoracic ATAC-Seq libraries relate to one another, principal component analysis was performed. There is some variation between libraries (PC2 = 16%), but most differences in the chromatin landscape are caused by a difference in axial level (PC1 = 49%) (Figure 4.2C). Read quality and alignment efficiency was high, with less than 30% of all reads displaying mitochondrial contaminants and low duplication rates. The changes across the chromatin landscape must be subtle and at specific loci as the correlation between all samples was high genome-wide (Figure 4.2A). These ATAC-Seq libraries have the characteristic visual profiles of ATAC-Seq libraries, with peaks denoting regions of accessible chromatin around each locus (Figure 4.2B). Biological replicates for each sample were merged to generate overall alignment files for each condition (cervical and thoracic). As expected, all ATAC-Seq libraries have a normal fragment length distribution, with the majority of sequenced reads either nucleosome-free or of mononucleosome-associated lengths (Buenrostro *et al.*, 2013). I conclude that all libraries are of good quality and can be used to assess the functional regulation of HOX genes across the cervical-thoracic axial boundary at the chromatin level.

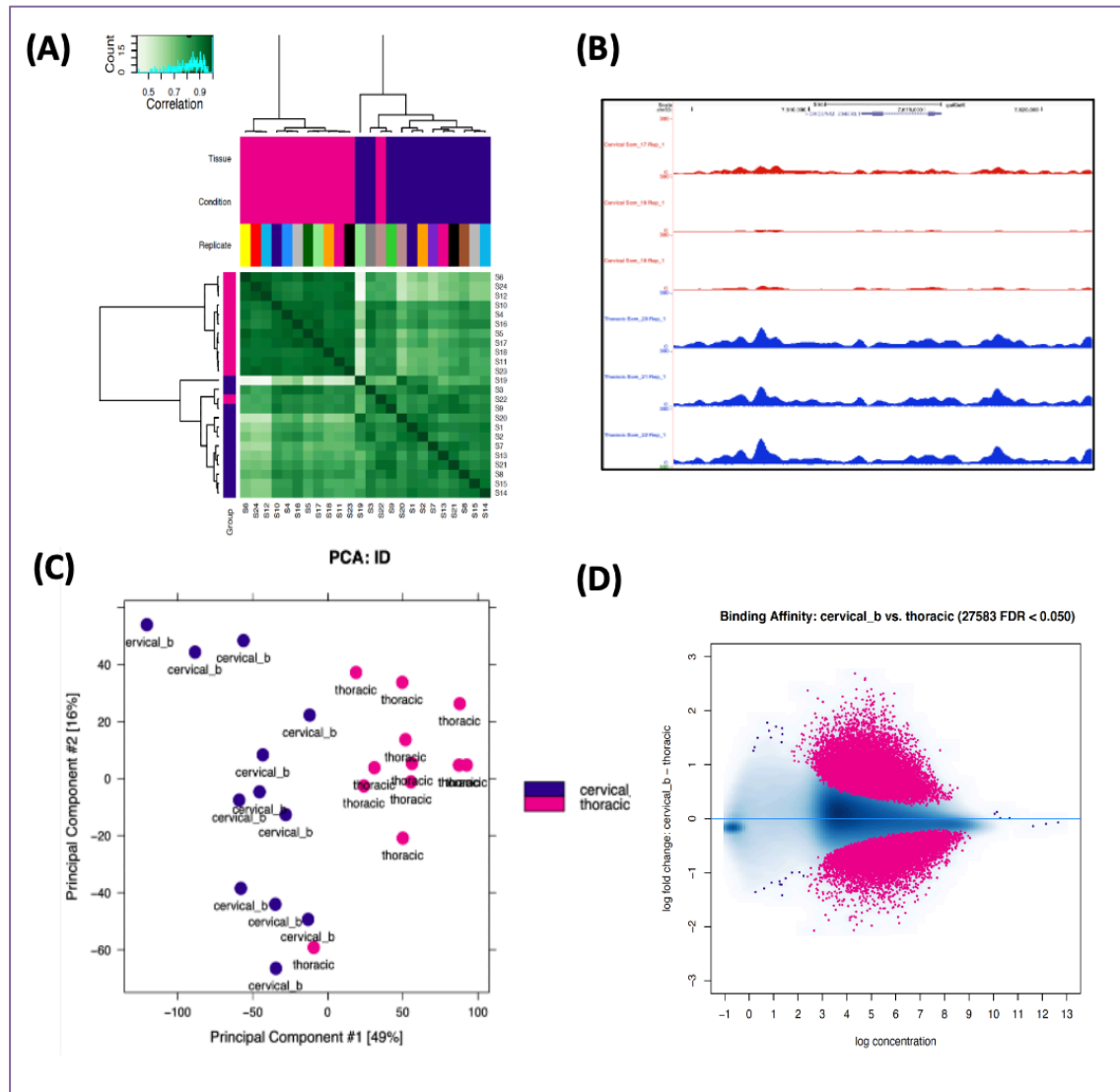


Figure 4.2: Quality control assessment of ATAC-Seq libraries. **(A)** Correlation heat map of accessible chromatin regions (ATAC-Seq peak sites) comparing cervical somites (purple) and thoracic somites (pink). **(B)** IGV track from the UCSC genome browser displaying chromatin accessibility within the individual somites from an example replicate for each somite number (17-22) around the HOXC8 locus. **(C)** Principal component analysis (PCA) generated using ggplot (version 2.2.1) within the R programming environment (version 3.3.1) for ATAC-Seq libraries generated in biological quadruplet. The conditions assessed are all replicates of cervical somites (purple) vs. thoracic somites (pink). **(D)** MA plot of significantly differential peak sites (pink) comparing cervical somites with thoracic somites.

4.2.2. Selection of Hox genes to investigate

In animals, Hox genes encode a family of transcription factors that are important for proper development of this axial formula. Hox gene types are related to segment type not segment number, e.g. in chicks cervical (neck) segments extend to somite 19, whereas in mammals they extend to somite 12. Yet regardless of the segment number, the change from cervical-type to thoracic-type is marked by the same transition of Hox gene expression (Fig 4.3A, B).

Hoxc6, a member of the sixth paralogue group, has an anterior boundary of expression at the level of the first thoracic vertebrae in both the chick and the mouse. In the chick, Hoxc6 labelling begins at somite level 19-20, two vertebrae behind Hoxc5 expression, and is also aligned with the middle of the forelimb bud. This has been confirmed using *in situ* hybridisation on a wild type stage HH16 chick embryo (Fig. 4.3C) In the mouse Hoxc6 expression begins at somite level 12-13 and is also two vertebrae behind Hoxc5 expression and in axial alignment with the middle of the forelimb bud (Burke *et al.*, 1995). Hoxc6 expression demarcates the anterior boundary of the thoracic region in many vertebrates including mouse and chick (Gaunt 1994; Burke *et al.*, 1995), alligator (Mansfield & Abzhanov, 2010) and snakes (Woltering *et al.*, 2009). Overexpression of Hox6 paralogues in PSM produces ectopic ribs in more anterior cervical and more posterior lumbar regions, indicating that Hox6 can determine thoracic identity throughout the vertebral column. Hoxc8 is another example of a Hox gene expressed throughout the thoracic domain with its expression occurring typically 2 somites posterior to the C-T boundary. I hypothesise that these three HOX genes are responsible for the transition from cervical to thoracic identity in somites and their gene regulation will be investigated throughout this chapter.

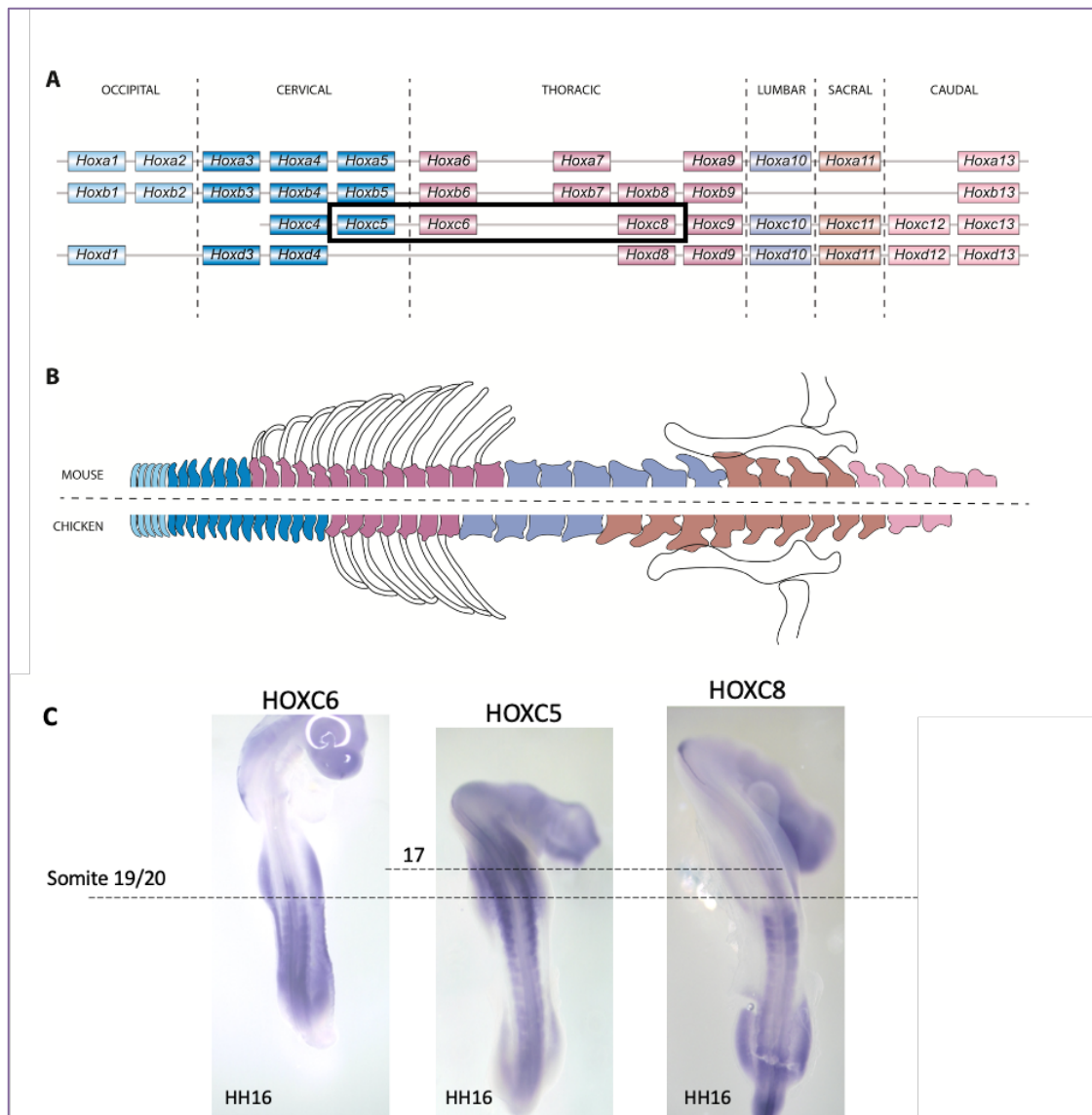


Figure 4.3: Hox gene expression pattern. **(A)** Schematic of the genomic organisation of the four Hox clusters (a–d). Hox gene combinatorial expression controls vertebral identity along the anteroposterior axis. Paralogue genes are colour-coded based on their contribution to axial domains (light blue = occipital; dark blue = cervical; dark pink = thoracic; purple = lumbar; brown = sacral; light pink = caudal). Boxes placed perpendicularly depict paralogous relationships within the Hox clusters. Genes of particular focus for this project are highlighted by the black box. **(B)** Schematic representation of the vertebral formula of the mouse compared to the chicken. The somite-derived vertebrae are aligned and subdivided into colour-coded domains corresponding to axial domains, which correlate with specific Hox paralogue expression (A). **(C)** *In situ* hybridisation of HOXC6, HOXC5 and HOXC8 in wild type chick embryos stage HH16. Black dotted line marks the transition from the cervical to thoracic boundary at somite level 19/20. An additional dotted line indicates somite no. 17, the anterior expression boundary of HOXC5. Somite 17 is the first of six individual somites used for RNA-Seq and ATAC-Seq in this project.

In order to establish that Hox expression is established early in development in the PSM prior to somite formation, and subsequently fixed following heterotopic graft of the tissue, an *in ovo* graft of one somite and the following PSM of thoracic tissue, equivalent to two somites in length, was transplanted to the cervical domain of a younger chick embryo. The early expression of HOXC6 and subsequent fixing of its expression in the tissue, adds strength to the hypothesis that this gene is responsible for the determination of the thoracic domain. The graft was carried out on the right side of the embryo only, leaving the left side as a control for comparison. Following whole mount *in situ* hybridisation, *HOXC6* is shown to be expressed ectopically in the grafted tissue, indicating that axial level identity was specified at the time of grafting (Fig. 4.4).

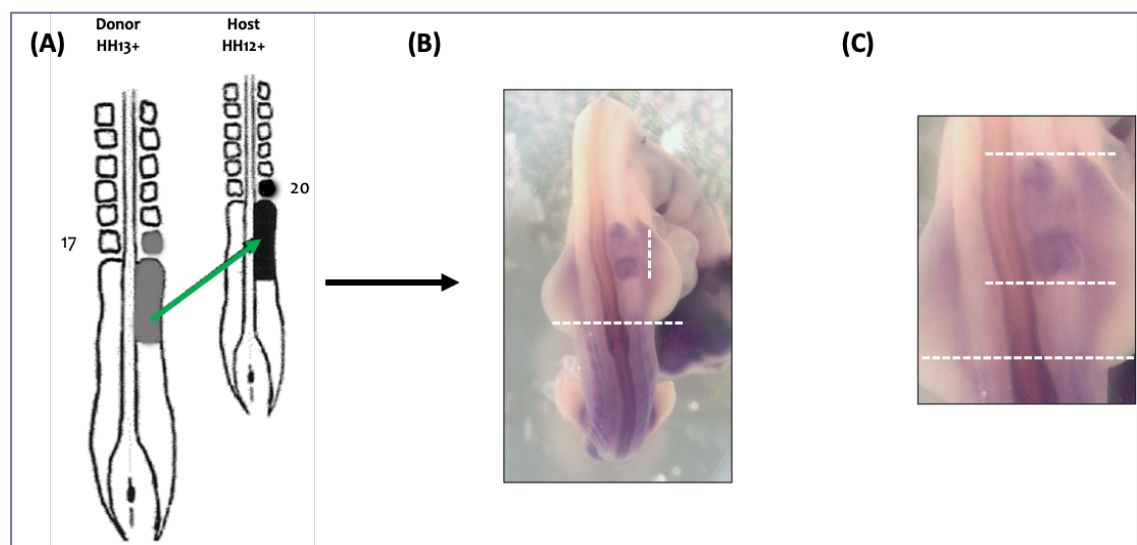


Figure 4.4: Ectopic *HOXC6* expression following thoracic to cervical graft. (A) Schematic of the posterior to anterior *in ovo* graft of one somite and the following PSM, equivalent to two somites in length, of thoracic tissue. (B) Ectopic *HOXC6* expression three days post graft. The boundary of *HOXC6* in thoracic level somites is indicated by the horizontal white dashed line. (C) Enhanced image of graft segment.

4.2.3. Identifying and validating Hox *cis*-regulatory elements

I initially analysed the expression profiles of the three HOXC genes of interest - *HOXC5*, *HOXC6* and *HOXC8* – in HH16 chick embryos in order to capture visually the recently formed cervical-thoracic boundary and confirm the location where expression of each Hox gene of interest begins across this boundary. Hox genes are generally expressed in the paraxial and lateral mesoderm of the main embryonic axis, as well as in the neural tube, and display progressively more restricted expression domains depending on their position within the cluster (Gaunt *et al.*, 1988). Comparative WISH analysis of chicken, mouse, and *Xenopus* embryos has shown that HOXC expression patterns in the trunk are markedly similar across vertebrate species, and

consistently mark the same boundary. Whole mount *in situ* hybridisation (WISH) confirmed that *HOXC5*, *HOXC6* and *HOXC8* are transcribed consecutively across the cervical-thoracic boundary. While these three genes have an overlapping function, with all three influencing the transition from cervical cell fate in somites to thoracic cell fate, all three HOXC genes are activated at different developmental times. The precise timing of transcriptional activation typical of Hox genes suggests that there may be a dependence on a specific set of unique CREs for each gene, which are different from the CREs controlling the neighbouring paralogues.

4.2.3.1. Establishing methodology

First, I examined open chromatin peaks flanking the *HOXC8*, *HOXC6* and *HOXC5* genes within the areas both upstream and downstream of the genes that did not contain the transcript for a different gene. Peaks identified represented candidate *cis*-acting regulatory elements (CREs) and were then cloned upstream of the herpes simplex virus thymidine kinase (HSV-TK) minimal promoter, driving expression of a stable Citrine reporter.

Attempts to experimentally validate putative enhancer elements, using the Citrine reporter constructs, were made following a similar electroporation method used regularly in our lab (Mok *et al.*, 2021). This method was adopted from (Williams *et al.*, 2019). Electroporation targeted the prospective mesoderm of gastrula-stage HH3 + embryos and reporter gene expression profiles were monitored until HH11 (Fig 4.5). However, despite this method successfully identifying many other putative enhancers, it was not possible to validate putative enhancer elements for the Hox genes using this approach. This is most likely due to the very late expression of the three Hox genes of interest: *HOXC8*, *HOXC5* and *HOXC6*. Anterior expression of these Hox genes begin around stage HH12, when the somites across C-T boundary are beginning to form and we would need embryos to survive to even older stages in order to visualise and confirm the pattern of expression for the putative enhancer element after electroporation. These later stages were difficult to obtain using this method as survival of embryos *ex ovo* diminished after stages HH11, with healthy growth of the embryo greatly decreasing once the head begins to turn on the EC culture plates. We can see this reduction in embryo quality below (Fig 4.5), where the older embryos look unhealthy, with some developmental defects occurring, such as lack of some physical features (e.g. somites) due to the physical barrier of the EC culture plate. Furthermore, the decrease in essential nutrients over time may contribute to disruption in embryo growth and decreased survival.

As well as this, the reporter construct becomes diluted due to cell proliferation over the 2-3 days of embryo development and growth to HH12-16. This results in decreased level of fluorescence that is not sufficient for detection in any healthy embryos that survived to a suitable stage.

The three putative enhancers to be tested, for HOXC5, HOXC6, and HOXC8, were not expected to show signal until at least stages HH12 and embryos would need to survive to even later stages in order to show optimum expression for validation. Therefore, the experimental procedure was later changed to *in ovo* electroporations targeting the PSM and neural tube on the right side of stage HH10 embryos. Citrine reporter gene expression profiles were monitored until HH16.

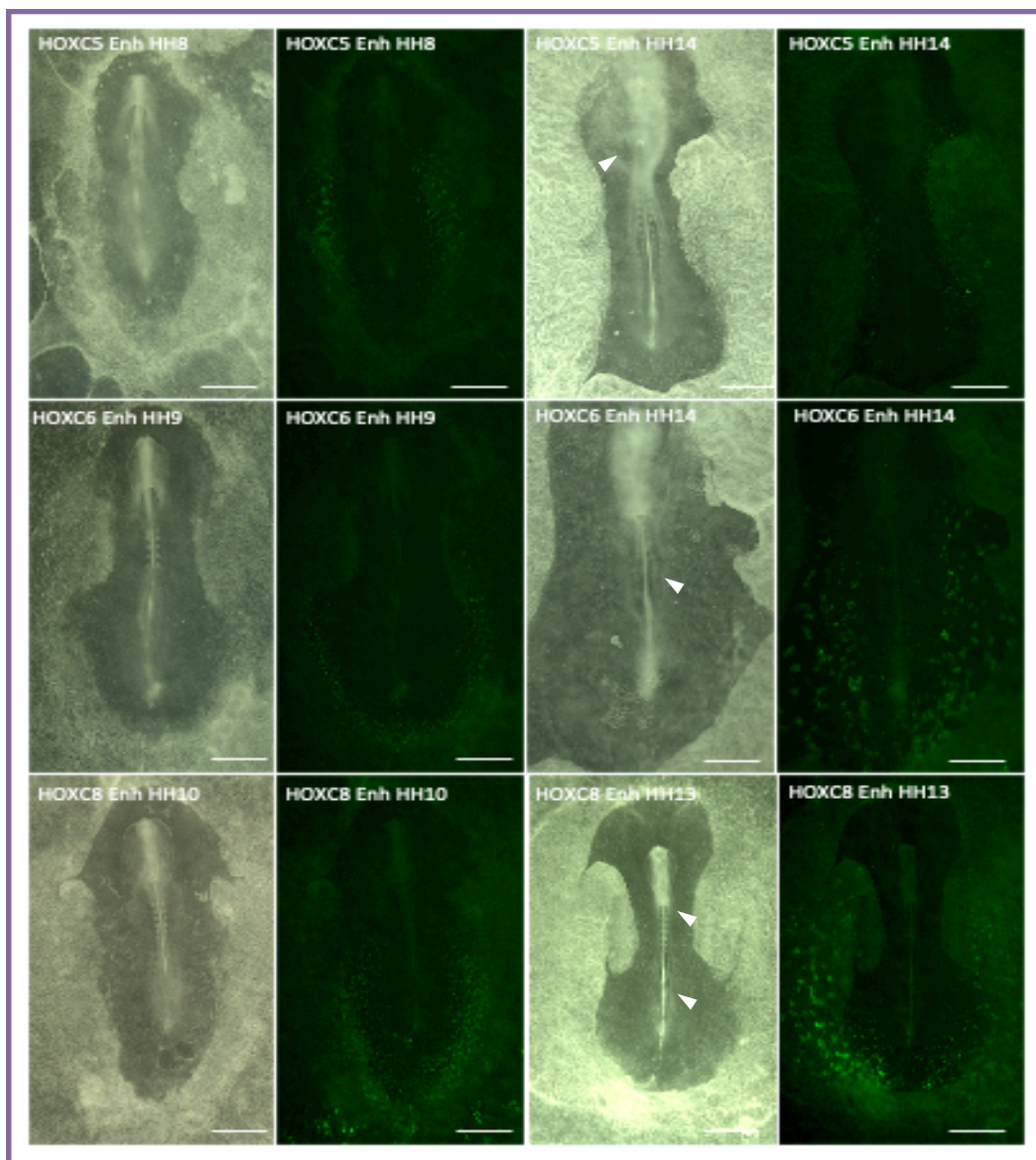


Figure 4.5: Initial attempts at candidate enhancer validation *ex ovo*. Images of EC cultured chick embryos 24hr (HH8-20) and 48hr (HH13-14) following injection and electroporation of three candidate enhancers for the genes *HOXC5* (n = 24), *HOXC6* (n = 38) and *HOXC8* (n = 43). Embryos were placed and imaged ventral side up. White arrows indicate areas where typical features of development are missing, such as a ballooning heart and somites.

4.2.3.2. Identification and validation of *HOXC8* enhancer

In order to determine that the method of selection for putative enhancer elements was effective, as a control I made use of a previously identified *HOXC8* enhancer element identified by deletion in mice and later proved to be conserved in chick (Shashikant *et al.*, 1995). This element was described as a 200bp fragment located 3kb upstream of *HOXC8*. To determine if I could locate and easily identify this already known enhancer element using the ATAC-Seq dataset only, I first looked for the gene of interest on the UCSC browser. *HOXC8* gene location is highlighted in yellow in the figure below (Fig 4.6A). Then, I hypothesised that by searching both upstream and downstream of that gene of interest for a differential peak that also shows high conservation with other vertebrates, I could identify putative cis-regulatory enhancer elements for the hox genes. The span for this search reached until the next nearest gene. The element highlighted in blue (Fig4.6A) showed both a differential peak, where chromatin was closed in the cervical somites (red peak) and open in the thoracic somite (blue peak), and had the highest level of conservation among the 77 vertebrate species accessible by the UCSC genome browser. After identifying this fragment as the most likely candidate for enhancer regulation of *HOXC8*, I compared its sequence to that of the previously identified element. Both sequences were the same, thus confirming I successfully identified the *HOXC8* enhancer element and my method for selection of putative enhancers was effective. Subsequently, I used the same approach to identify regulatory elements for both *HOXC6* and *HOXC5*.

After successfully identifying the enhancer element for *HOXC8* using the ATAC-Seq data, I then needed to confirm if I could validate and thus further characterise these enhancer elements experimentally by embryo electroporation (Chapman *et al.*, 2001; Mok *et al.*, 2021). *In ovo* electroporation intended to target the presomitic mesoderm and neural tube of stage HH10 embryos and reporter gene expression profiles were monitored after 24hrs (Fig 4.7A). However, due to human error and a lack of time necessary to improve targeting of electroporation, only the neural tube was targeted. This was confirmed using an mCherry construct that is expressed in all cells and which was successfully electroporated. Neural tube electroporation was still adequate for determining if a candidate element displays enhancer activity as the Hox genes are also expressed there (Fig 4.3). The 200bp *HOXC8* element displayed enhancer activity in the neural tube (Fig 4.7A). Citrine reporter expression was compared to *HOXC8 in situ* expression (Fig 4.3). The anterior boundary of citrine expression in the neural tube was found to reflect the expression of the endogenous *HOXC8*. Expression of the ubiquitous mCherry construct by comparison extended more anteriorly to the citrine reporter, which highlighted the location across the neural tube where enhancer activity for *HOXC8* is activated.

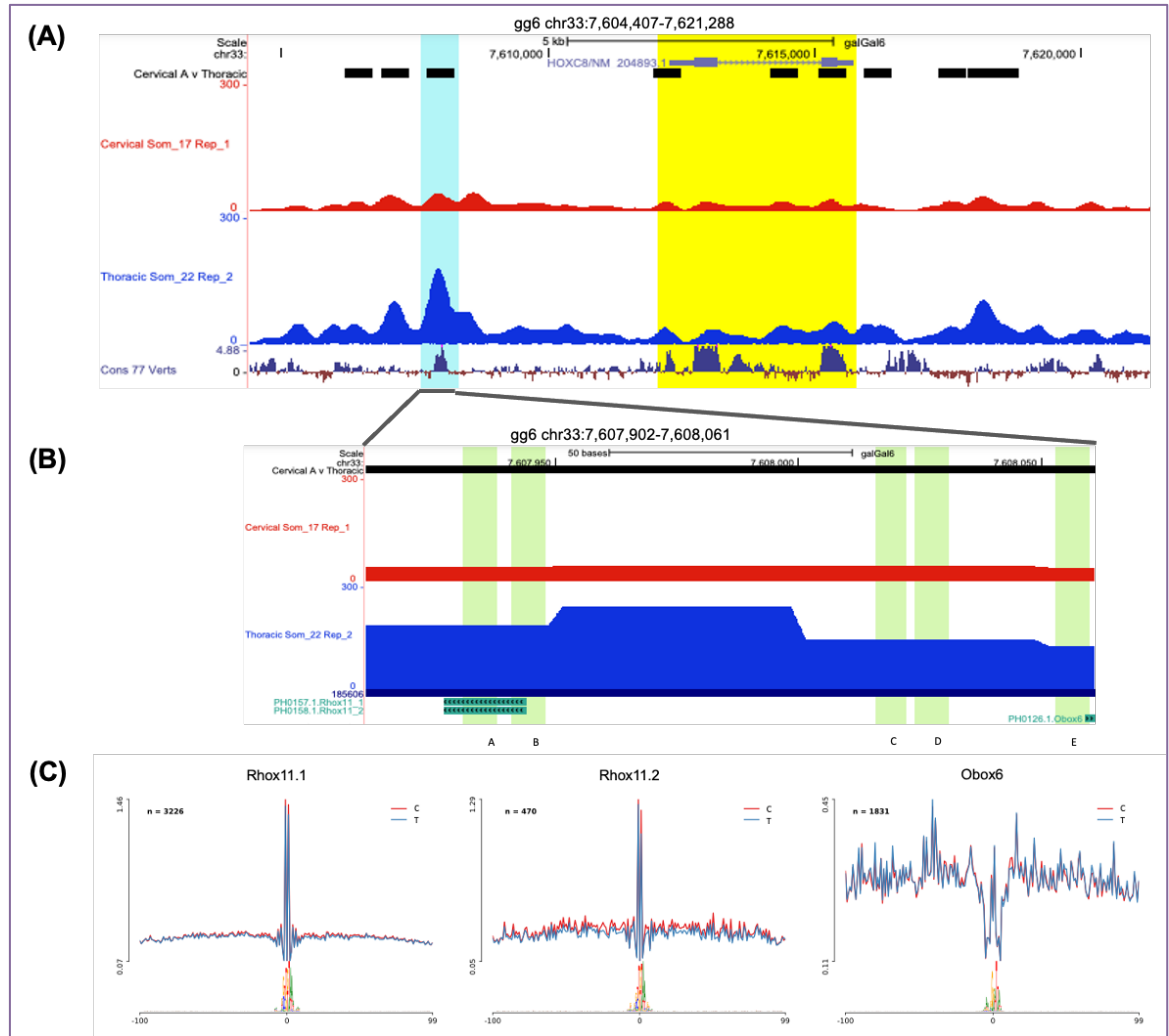


Figure 4.6: Identification of a regulatory element for *HOXC8*. (A) ATAC-Seq profile at the *HOXC8* locus. Yellow box indicates the *HOXC8* gene. Blue box indicates the putative enhancer identified. Black bars indicate location for all differential peaks identified across the *HOXC8* locus. (B) Enhanced image of putative enhancer with five 'zones of activity' highlighted in light green. Differential footprints identified across the five zones of activation indicated by green bars (labelled A-E). (C) Coverage of Rhox11.1, Rhox11.2 and Obox6 footprints across the cervical (red) and thoracic (blue) somites identified within zones of activity A, B and E of the putative *HOXC8* enhancer element.

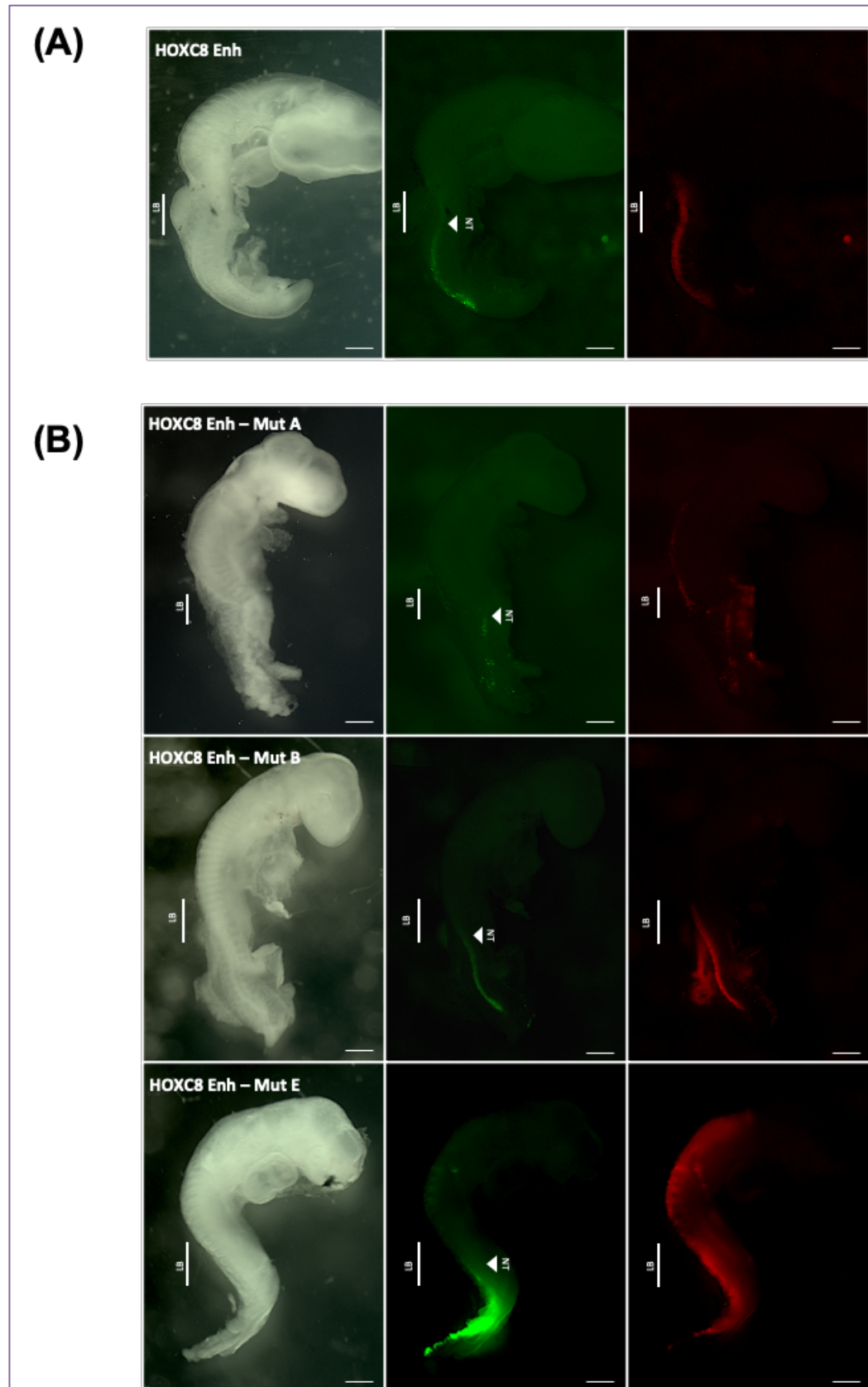


Figure 4.7: Visualisation of a regulatory element for *HOXC8*. **(A)** *HOXC8* enhancer (n = 13/13) citrine reporter expression in neural tube (NT). **(B)** *HOXC8* enhancer (n = 2/3) citrine reporter expression in neural tube (NT) following mutagenesis of 'zone of activity' A; *HOXC8* enhancer (n = 4/4) citrine reporter expression in neural tube (NT) following mutagenesis of 'zone of activity' B; and *HOXC8* enhancer (n = 4/4) citrine reporter expression in neural tube (NT) following mutagenesis of 'zone of activity' E.

To further interrogate the accessible chromatin landscape during somite development, we used HINT-ATAC (Li *et al.*, 2019) to discover differential TF footprints in regions of open chromatin identified between cervical somites and thoracic somites. I identified three TF footprints within the enhancer, one for RHOX11.1, one for RHOX11.2, and one for Obox6 (Fig 4.6B). Previous studies have carried out mutational analysis within this 200bp fragment and revealed five distinct *cis*-acting elements (labelled A-E) that are potential binding sites for transcription factors. These are very short elements, all just 7bp in length. Some of these elements were found to be redundant whereas others were interdependent (Shashikant & Ruddle, 1996). These five elements are highlighted in green and labelled A-E above (Fig 4.6B). However, no specific transcription factor has been identified as necessary for *HOXC8* enhancer expression. The TFs that may bind within these five elements are speculative and have yet to be confirmed. With the use of HINT-ATAC, I was able to narrow down the possible elements important for *HOXC8* enhancer activation across the C-T boundary to just three of the five elements: A, B, and E (Appendix 2). Two of these elements (A and B) share the same TF footprint site for RHOX11, possibly indicating that this is in fact one extended *cis*-acting element rather than two separate elements as previously thought.

By mutating the three individual elements: A, B, and E, I next determined their requirement for the activation of fluorescent reporter expression. Mutation of the element E, which overlapped with the Obox6 footprint, had no effect and strong reporter activity was still observed following *in ovo* electroporation (Fig 4.7B). Mutation of elements A and B individually led to a significant decrease in reporter activity. The current mutation of individual sites only partially captures the length of the RHOX11.1/RHOX11.2 TF footprint. I hypothesise that mutation of both elements would result in a complete loss of reporter activity as the entire TF binding site would be removed. Unfortunately, due to time constraints I was unable to test this idea.

As both RHOX11.1 and RHOX11.2 were identified to have a binding site across elements A and B, I next determined which specific TF may be required for the activation of this enhancer. HINT-ATAC identified a greater number of RHOX11.2 footprints in open chromatin in the cervical region when compared to the thoracic region (Fig 4.6C), suggesting that this TF may mediate the change in expression of the *HOXC8* enhancer element across the C-T boundary. In contrast, *Rhox11.1* coverage is not differential across the boundary. *OBOX6*, which overlaps element E and extends beyond the length of the enhancer fragment and does not show a significant change in the number of footprints across the C-T boundary. This is in agreement with the observation that there was no decrease in fluorescence following mutation of element E, suggesting *OBOX6* does not play a role in regulating *HOXC8* enhancer activity across the C-T boundary.

Although RHOX11.2 is the only transcription factor highlighted as differential across the C-T boundary within this enhancer element, not much is known about its function within the chick. RHOX genes are a family of homeobox genes clustered on the X chromosome in mice. They are preferentially expressed in reproductive tissues and are thus thought to contribute to mammalian fertility. Similar to other homeobox genes, the RHOX genes are expressed during embryonic development but there is not a lot of information available about their role during this time, particularly for RHOX11. Several members of the RHOX cluster remain expressed at high levels after birth and thus are candidates to control postnatal and adult developmental events (MacLean, 2013). Because of this, research into this gene has primarily been restrained to much later adult stages of mice development rather than their roles in early embryo development. Research has also been restricted to mammals, particularly mice and rats. Quantitative trait locus (QTL) analysis, a statistical method that links phenotypic data (trait measurements) and genotypic data (usually molecular markers), in an attempt to explain the genetic basis of variation in complex traits, has been carried out in rats. This indicated a possible link between RHOX11 and bone structure and strength as well as joint integrity, indicating a much wider role for RHOX11 than previously researched.

4.2.3.3. Identification and validation of HOXC6 enhancer

Using the same method of identification for the putative *HOXC8* enhancer element, I identified another enhancer element, which is 336bp long and located 12kb upstream of *HOXC6* (Fig 4.8A). This element also displayed enhancer activity in the neural tube following electroporation (Fig 4.9A). The expression pattern was compared to that of the *HOXC6 in situ* expression (Fig 4.23) and the anterior boundary for neural tube expression was found to reflect the expression of the endogenous *HOXC6* at somite level 17. Expression of the mCherry construct by comparison extended more anterior to the citrine enhancer reporter, highlighting the exact location across the neural tube where enhancer activity for *HOXC6* is activated. Several additional differential peaks were also cloned and tested around the *HOXC6* locus, but no enhancer activity was detected from those fragments (data not shown) (Appendix 1).

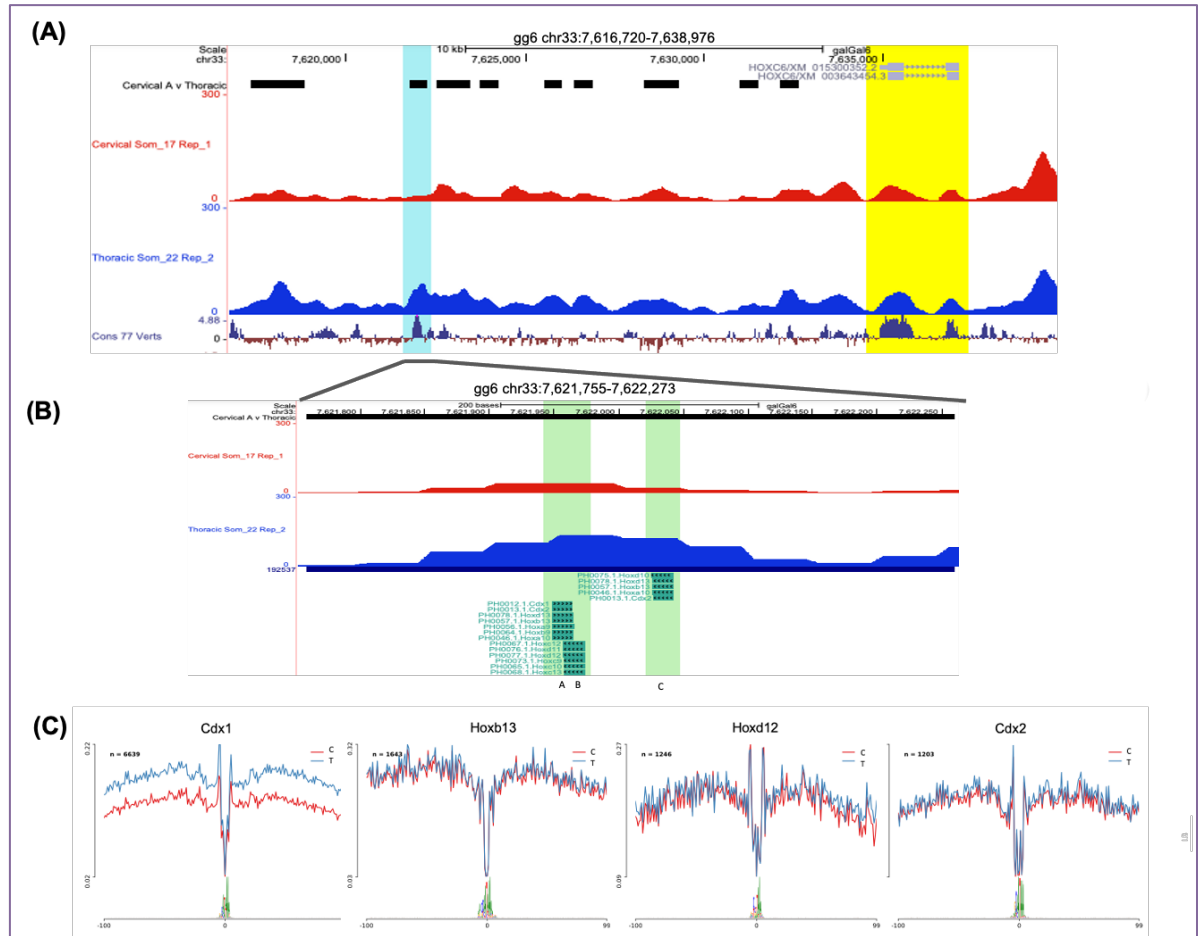


Figure 4.8: Identification of a regulatory element for *HOXC6*. **(A)** ATAC-Seq profile at the *HOXC6* locus. Yellow box indicates the *HOXC6* gene. Blue box indicates the putative enhancer identified. Black bars indicate location for all differential peaks identified in this region. **(B)** Enhanced image of putative enhancer with three sites for potential transcription factor (TF) footprint mutagenesis highlighted in light green. Differential footprints identified across the three zones of activation indicated by green bars (labelled A-C). **(C)** Coverage of CDX1, HOXB13, HOXD12 and CDX2 footprints across the cervical (red) and thoracic (blue) somites.

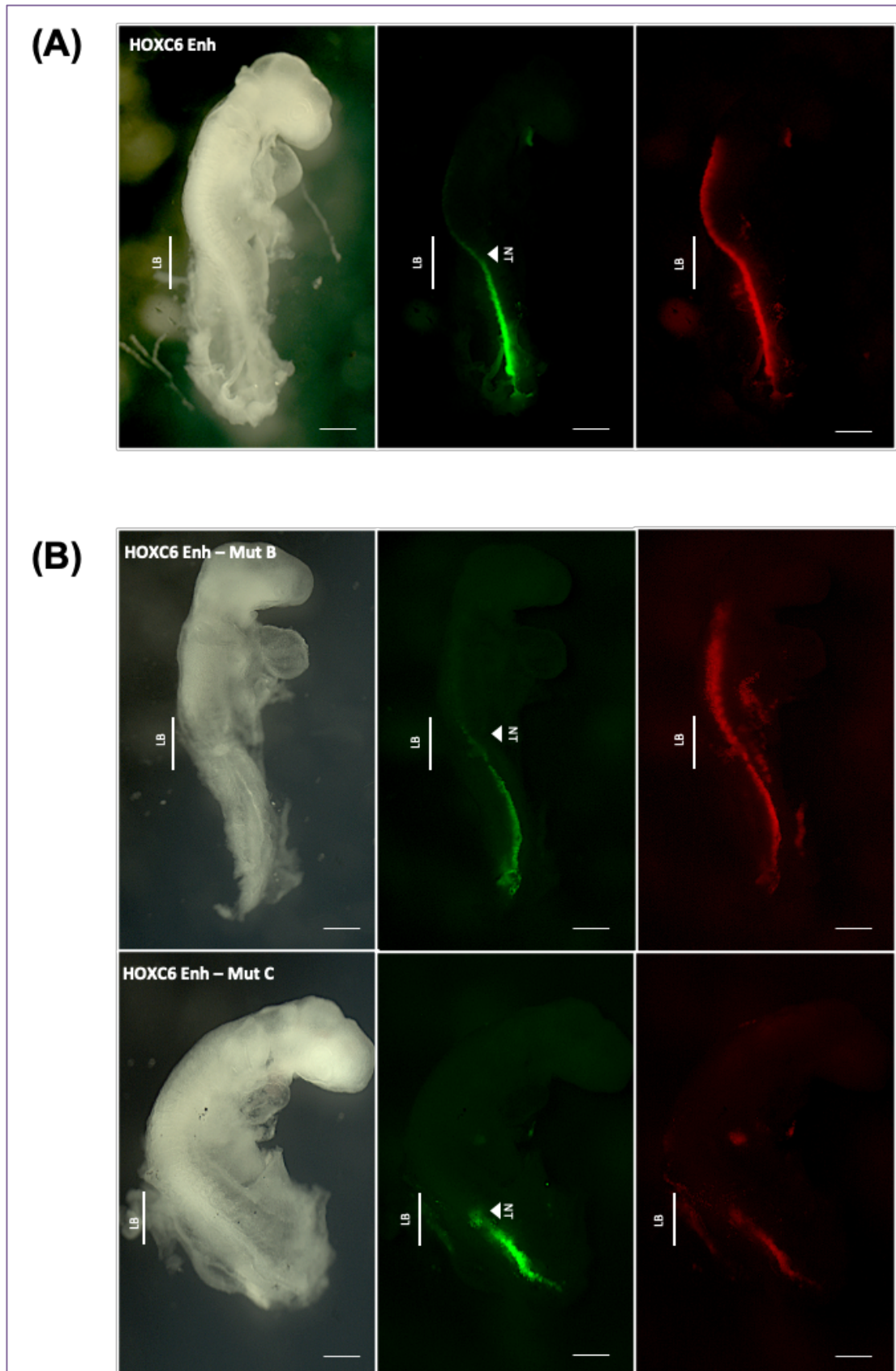


Figure 4.9: Visualisation of a regulatory element for *HOXC6*. **(A)** *HOXC6* enhancer (n = 10/10) citrine reporter expression in neural tube (NT). **(B)** *HOXC6* enhancer (n = 3/3) citrine reporter expression in neural tube (NT) following mutagenesis of TF binding site B and *HOXC6* enhancer (n = 3/3) citrine reporter expression in neural tube (NT) following mutagenesis of TF binding site E.

HINT-ATAC (Li *et al.*, 2019) again discovered differential TF footprints in regions of differentially open chromatin identified between cervical somites and thoracic somites. I identified three regions of interest (A-C) within the enhancer, with two regions slightly overlapping (Fig 4.8B). All three footprints were 16bp long. I next determined their requirement for the activation of fluorescent reporter expression by mutating individual elements. Mutation of the footprint sites B and C individually had no effect, following *in ovo* electroporation strong reporter activity was still observed (Fig 4.9B). I hypothesise, that a mutation of footprint site A may affect reporter activity and that the transcription factor that binds there could be critical for *HOXC6* enhancer activity. However, due to time constraints I was unable to complete the mutagenesis for footprint A.

I next determined which specific TF may be required for the activation of the putative *HOXC6* enhancer. All three sites share a similar list of possible TFs that could bind there. On closer inspection of the footprints that are called within sites B and C, for example CDX1, HOXB13 and HOXD12 (Fig 4.8C), none of these show a greater number of footprints in either the cervical or the thoracic somites. This is consistent with the finding that mutagenesis of those fragments did not lead to a decrease/change in citrine reporter expression. For site A, multiple HOX genes as well as CDX2 are identified as having a possible binding site within the *HOXC6* enhancer. I propose that these could be critical for *HOXC6* enhancer activity. However, the Hox genes listed are expressed much later and more posterior to the boundary captured, therefore they are unlikely to be expressed sufficiently to bind to the enhancer element. It is possible that these TFs may act as repressors of this enhancer more posteriorly. The two remaining footprints within site A are CDX1 and CDX2, both of which have been well documented to be involved in anteroposterior patterning and posterior axis elongation (van den Akker *et al.*, 2002). HINT-ATAC identified a greater number of CDX2 footprints in open chromatin in the thoracic region when compared to the cervical region (Fig 4.8C), indicating that this TF may have a role in the change in expression of the *HOXC6* enhancer element across specifically the C-T boundary. In contrast, CDX1 is not significantly differential across the boundary. Interestingly, CDX2 was not called as a differential footprint within the *HOXC8* enhancer across the C-T boundary. Alongside expression pattern, this indicates that it is primarily *HOXC6* that is responsible for the sudden change in identity from cervical somite to thoracic somite. *HOXC8* may be responsible for more downstream effects or effects in neural tissue, which would not have been picked up in our examination of somites only.

4.2.3.4. Identification and validation of *HOXC5* enhancer

Using the same method of identification for the putative *HOXC8* and *HOXC6* enhancer elements, I identified another enhancer element, 658bp long and located 13kb downstream of *HOXC5* (Fig 4.10A). This element also displayed enhancer activity in the neural tube following electroporation (Fig 4.10B). The expression pattern was compared to that of the *HOXC5* *in situ* expression (Fig 4.3) and the anterior boundary for neural tube expression was found to reflect the expression of the endogenous *HOXC5* expression pattern seen in whole mount *in situ*. Expression of the mCherry construct extended more anteriorly to the citrine enhancer reporter, highlighting the exact location across the neural tube where enhancer activity for *HOXC5* is activated. Several additional differential peaks were also cloned and tested around the *HOXC5* locus, but no enhancer activity was detected from those fragments (data not shown) (Appendix 1).

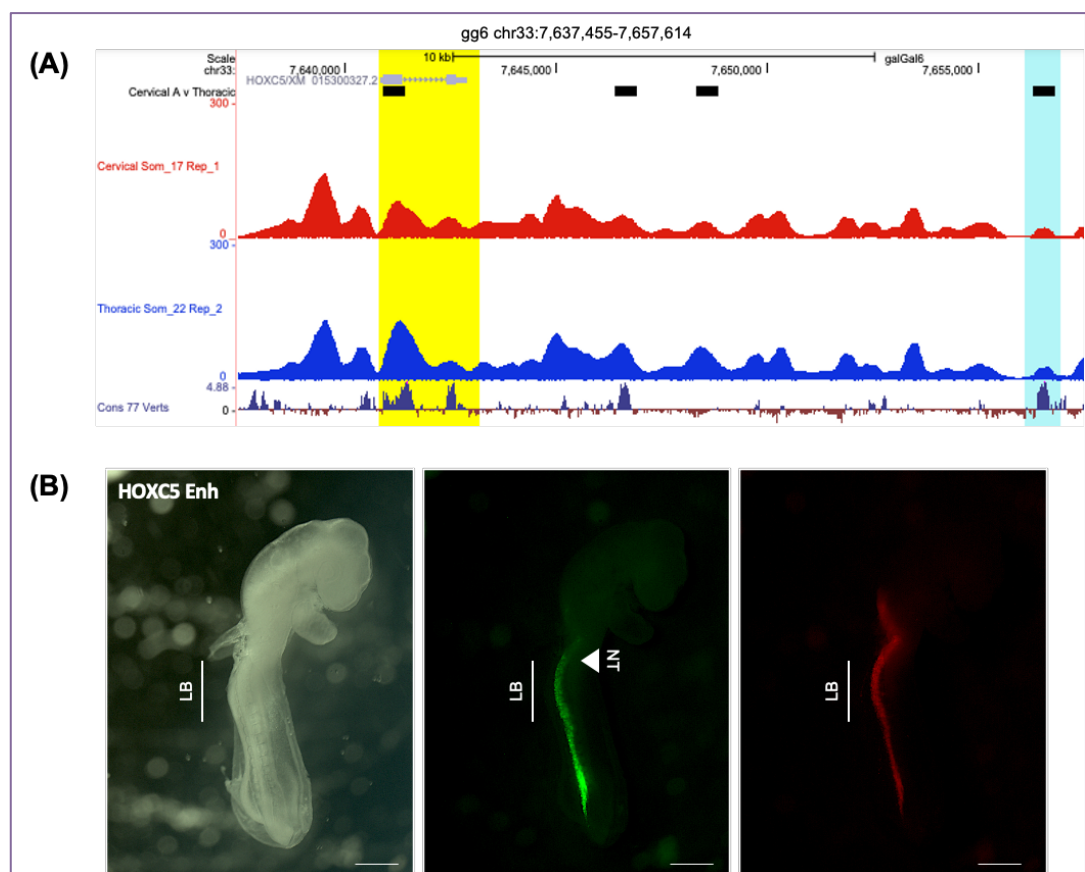


Figure 4.10: Identification of a regulatory element for *HOXC5*. (A) ATAC-Seq profile at the *HOXC5* locus. Yellow box indicates the *HOXC5* gene. Blue box indicates the putative enhancer identified. Black bars indicate location for all differential peaks identified across the *HOXC5* locus. (B) *HOXC5* enhancer (n = 8/8) citrine reporter expression in neural tube (NT).

No differential footprints were identified within the *HOXC5* enhancer elements. This was to be expected as all somites investigated across the C-T boundary would express *HOXC5*, which is expressed two somites anterior to the boundary. Thus, any footprints would not necessarily be called as differential. This also explains the very subtle changes in chromatin accessibility, with both cervical and thoracic somites shown to have a similar level of openness visually in this region (Fig 4.10A). Despite the subtle change in chromatin accessibility indicating a differential peak, there were not enough differences between cervical somites versus the thoracic somites to be called as differential.

4.2.3.5. Genome wide search for Hox cluster enhancers

In order to utilize this large dataset to a wider degree, as well as to showcase broader utility of the dataset rather than just the niche examination across the C-T boundary, I carried out a wide scale search for putative enhancers across the remaining three Hox clusters. Imaged below are the genome browser views of the three remaining Hox clusters (A, B and D), with the Hox gene marking the beginning and end highlighted in yellow (Fig 4.11A-C). A few candidate *cis*-regulatory elements from each cluster were successfully cloned and tested (highlighted in blue). Several additional differential peaks or highly conserved fragments were not cloned successfully so therefore not tested (data not shown) (Appendix 1). Putative enhancers that were successfully identified are highlighted in green (Fig. 4.11A, B, C).

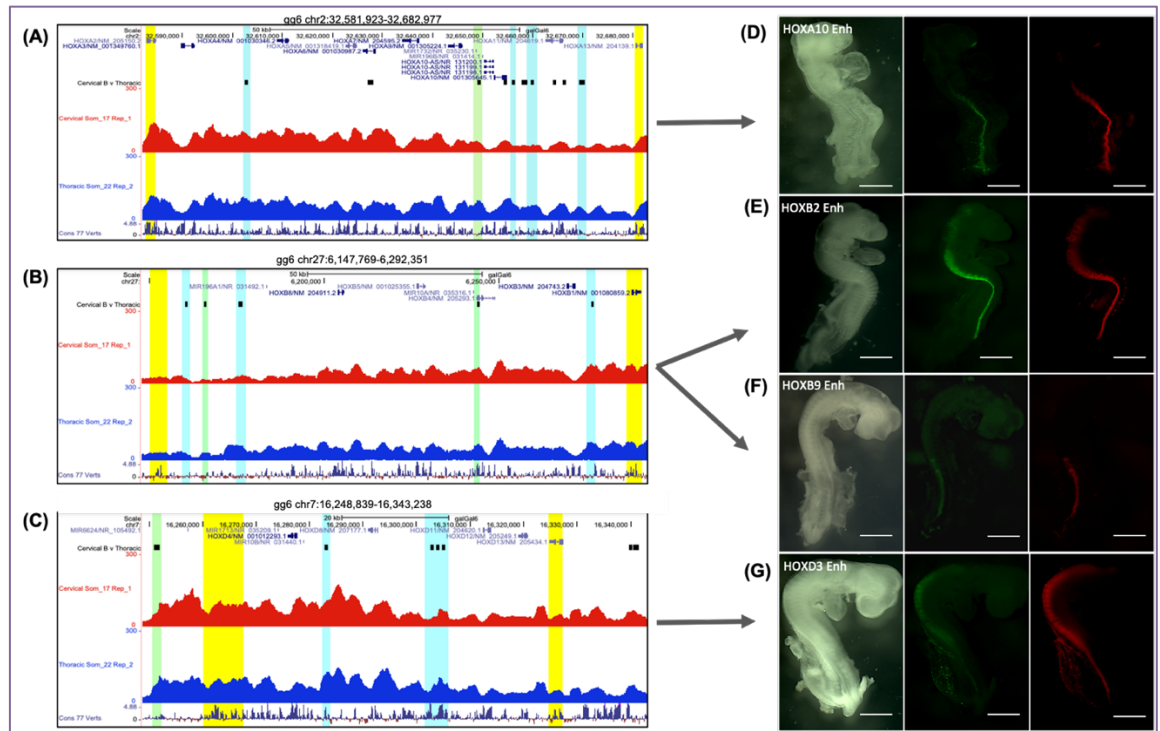


Figure 4.11: Identification and validation of putative enhancers across Hox clusters A, B, and D. ATAC-Seq profile across the HOXA (A), HOXB (B), and HOXD (C) clusters. Yellow boxes indicate the first and last Hox gene of the cluster. Blue box indicates the candidate enhancers successfully cloned and tested. Green box indicates enhancers that showed activity. Black bars indicate location for all differential peaks identified across the HOXA locus. *HOXA10* enhancer (n = 3/3) (D), *HOXB2* enhancer (n=3/3) (E), *HOXB9* enhancer (n = 2/2) (F), and *HOXD3* enhancer (n = 4/4) (G) citrine reporter expression in neural tube.

Within Hox cluster A, I identified another enhancer element for *HOXA10*, 660bp long and located 1kb upstream of *HOXA10* (chr2:32,648,959-32,649,618) (Fig 4.11A). This element also displayed enhancer activity in the neural tube following electroporation (Fig 4.11D). The expression pattern was compared to that of the *HOXA10 in situ* expression (using the GEISHA database) to confirm expression along the neural tube.

Within Hox cluster B, I identified two enhancer elements, one for *HOXB2*, 758bp long and located 1kb upstream of *HOXB2* (chr27:6,276,163-6,276,920) and one for *HOXB9*, 779bp long and located 23kb upstream of *HOXB9* (chr2:32,648,959-32,649,618) (Fig 4.11B). Both elements displayed enhancer activity individually in the neural tube following electroporation (Fig 4.11E; Fig. 4.11F). Their expression patterns were compared to that of the corresponding genes *in situ* expression (using the GEISHA database) and the anterior boundary for neural tube expression was found to reflect the expression of the endogenous Hox expression, *HOXB2* and *HOXB9* respectively. Expression of the mCherry construct extended more anteriorly to the citrine enhancer reporters, highlighting the exact location across the neural tube where enhancer activities for *HOXB2* and *HOXB9* were activated. Unfortunately, the DNA preparation and

isolation of the *HOXB9* plasmid construct yielded a very low concentration, so I was unable to inject the same concentration as with the other plasmids. This resulted in a much lower level of fluorescence, but citrine expression was still visible and comparable with the mCherry plasmid, thus can still be included as a successful result. For the *HOXB2* citrine reporter, a few posterior somites along the right-hand side of the embryo were electroporated (Fig 4.11E). Within these somites mCherry was expressed strongly, whereas citrine was not detected in this tissue. This is to be expected for a *HOXB2* enhancer element, as *HOXB2* is not expressed within the somites.

Finally, within Hox cluster D, I identified another enhancer element, 961bp long and located 8kb upstream of *HOXD3* (chr7:16,251,031-16,251,991) (Fig 4.11C). This element also displayed enhancer activity in the neural tube following electroporation (Fig 4.11G). The expression pattern was compared to that of the *HOXD3 in situ* expression (using the GEISHA database) to confirm expression along the neural tube.

4.2.4. Identifying downstream targets of HOXC6

To identify further genes that may be involved in mediating the transition from cervical to thoracic identity in somites, potentially downstream of the *HOXC* genes of interest I correlated the list of differentially expressed genes from the RNA-Seq dataset generated in Chapter 3 with the list of TFs associated with differential footprints across the C-T boundary genome wide. After candidate genes were identified by the approach, I next examined whether differentially accessible chromatin (potential *cis*-regulatory elements), either upstream or downstream of the genes of interest, contained a *HOXC6* footprint. Such a footprint would suggest an interaction between these genes and *HOXC6*. A list of 9 genes in total were identified as potential downstream targets of *HOXC6* following its activation across the C-T boundary (Fig. 4.12A).

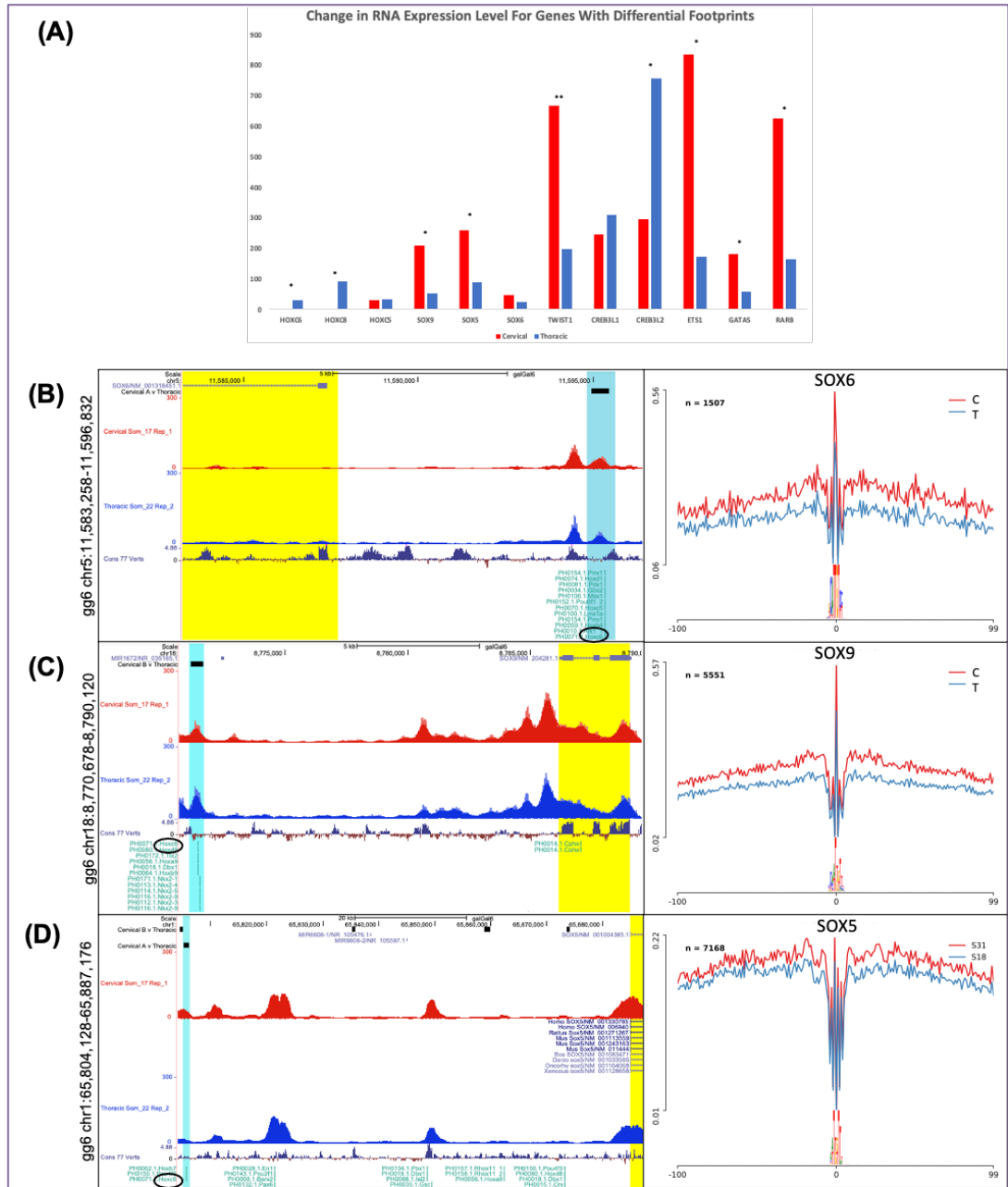


Figure 4.12: A selection of candidate targets of *HOXC6*. (A) Bar chart from the RNA-Seq dataset highlighting the change in RNA expression level of selected genes that showed both a change in RNA expression as well as a change in number of footprints when comparing cervical somites (red) to thoracic somites (blue). (B) ATAC-Seq profile upstream of the *SOX6* locus. Yellow box indicates the *SOX6* gene. Blue box indicates a potential *cis*-regulatory element containing a *HOXC6* footprint (black circle). The right-hand image shows the Tn5 insertion frequency across the C-T boundary containing at least one *SOX6* motif. (C) ATAC-Seq profile across the *SOX9* locus. Yellow box indicates the *SOX9* gene. Blue box indicates a potential *cis*-regulatory element containing a *HOXC6* footprint (black circle). The right-hand image shows the Tn5 insertion frequency across the C-T boundary containing at least one *SOX9* motif. (D) ATAC-Seq profile upstream of the *SOX5* locus. Yellow box indicates the *SOX5* gene. Blue box indicates a potential *cis*-regulatory element containing a *HOXC6* footprint (black circle). The right-hand image shows the Tn5 insertion frequency across the C-T boundary containing at least one *SOX5* motif.

A subset of genes identified as differential by RNA-Seq with a *HOXC6* footprint within a putative *cis*-regulatory element are three members of the SOX family: *SOX5*, *SOX6*, and *SOX9* (Fig. 4.12B-D). These three genes are often referred to as the ‘chondrogenic trio’. *SOX9* is a transcriptional activator required for chondrogenesis, and *SOX5* and *SOX6* are closely related DNA-binding proteins that critically enhance its function (Liu & Lefebvre, 2015). Due to their roles in chondrogenesis it is plausible to speculate that these genes may have a critical role implementing the change in vertebrae anatomy across the C-T boundary. However, not much is known about their function so early in embryonic development. Further studies, such as very early knockdown/overexpression of the Sox trio is required to confirm their function in CT somite differentiation.

To confirm that these genes are downstream targets of *HOXC6* and involved in mediating the transition from cervical to thoracic expression in somites, future experimentation is needed. RT-qPCR may be used to assess their expression in wild type somites and in somites where *HOXC6* is expressed ectopically. Should an increase in *HOXC6* expression result in a change in expression for any/all the 9 candidate genes identified, this would confirm interactions between these genes and further expand our knowledge on how somites commit to their cell fate across a fixed boundary and what genes are critical for these changes.

4.3. Conclusion

In this chapter, putative *cis*-regulatory elements, or enhancers, have been identified for the three key Hox genes hypothesised to have a major role in determining the change from somites fated to develop into cervical vertebrae to somites fated to develop into thoracic vertebrae – *HOXC5*, *HOXC6* and *HOXC8* - using an ATAC-Seq dataset generated from a line of 6 individual somites removed across the cervical-thoracic boundary. Before generation of the ATAC-Seq dataset, *in situ* hybridisation was carried out on wild type embryos to confirm the expected expression pattern of the HOX genes of interest across the C-T boundary. ATAC-Seq libraries were then made, and a series of quality control tests confirmed an accurate dataset was generated and that a significant difference could be seen across the C-T boundary. Enhancers were identified by locating differentially accessible peaks indicating sites of open chromatin both upstream and downstream of the genes of interest. Once a peak of interest had been identified, the candidate sequence was checked for conservation amongst vertebrates. Peaks with the largest difference in peak height and the most conservation amongst vertebrates were selected for cloning and electroporation.

Putative enhancers were identified for all three Hox genes of interest. For *HOXC8*, the gene was expressed two somites posterior to the cervical-thoracic boundary in all vertebrates, a differential peak was identified upstream of the gene. This differential peak marks the same area identified in mice as the early *HOXC8* enhancer by Shashikant *et al.* (1995). This enhancer acted as a control for our dataset, proving we could identify enhancers from our dataset while also establishing *in ovo* electroporation to validate those enhancers. Similarly, for *HOXC6*, the gene that marks the boundary across the cervical-thoracic domains across all vertebrates, a peak was identified upstream of the gene. For *HOXC5*, the gene expressed two somites anterior to the cervical-thoracic boundary in all vertebrates, a differential peak was identified downstream of the gene. In each case, chromatin accessibility correlated well with gene expression from the RNA-Seq dataset (see Chapter 3). The putative enhancers were then validated and further characterised by *in ovo* embryo electroporation.

Both *HOXC8* and *HOXC6* loci were further characterised using HINT-ATAC to select differential footprints across the candidate enhancer sequences. This identified transcription factors that occupy sites within the enhancers, and thus could be responsible for its activity. Mutagenesis of these sites identified RHOX11 as a key TF binding upstream of *HOXC8* and CDX1 as the key TF binding to the *HOXC6* enhancer. CDX1 is known to have a critical role in axial elongation and anteroposterior patterning in the embryo. This TF, along with CDX2, only shows as differential within the *HOXC6* enhancer and not the *HOXC8* enhancer. Therefore, I hypothesise that the key gene ultimately responsible for triggering and determining the change in domain identity from cervical to thoracic within somites is *HOXC6*.

Putative enhancers were identified for multiple genes across all four of the Hox clusters due to the extensive information ATAC sequencing provides. Candidate CREs for *HOXA10*, *HOXB2*, *HOXB9*, and *HOXD3* have been identified and validated by *in ovo* electroporation.

To identify further downstream genes that may be involved in mediating the transition from cervical to thoracic identity in somites, I correlated the list of differentially expressed genes from the RNA-Seq dataset generated in Chapter 3 with the list of differentially expressed footprints across the C-T boundary. A total of 9 potential downstream targets of *HOXC6* were identified following detection of a differential *HOXC6* footprint within potential *cis*-regulatory elements for each of those genes (Appendix 2). Future examination by increasing *HOXC6* expression and using RTqPCR to quantify changes of the 9 candidate target genes is required to further validate these interactions experimentally.

CHAPTER 5: LONG NON-CODING RNAs

5.1. Introduction

Long non-coding RNAs (lncRNAs) are often defined as transcripts >200nts long that have no discernible protein-coding ability (Quinn and Chang, 2016). Although relatively little is understood about the molecular mechanisms of lncRNA function, they have established roles in regulation of gene expression during development, cell differentiation and pluripotency (Fatica and Bozzoni, 2014; Luo *et al.*, 2016; Quinn and Chang, 2016; Rin and Chang, 2012) across diverse organisms ranging from plants to humans (Ulitsky and Bartel, 2013). lncRNAs have also been associated with numerous pathological conditions, such as cancers (Brunner *et al.*, 2012), cardiovascular disease and neurodegeneration (Chen *et al.*, 2013). A genomic region that has revealed several functional lncRNAs in diverse organisms are the Hox complexes (Pauli *et al.*, 2011; Pettini, 2012; Rinn *et al.*, 2007). The Hox complexes encode a set of transcription factors (TFs), physically clustered in the genome, which provide morphological identity along the anterior to posterior axis of developing embryos (Mallo and Alonso, 2013), throughout the majority of bilaterian animals (Moreno *et al.*, 2011). Misexpression or mutation of Hox genes causes morphological and pathophysiological defects (Quinonez and Innis, 2014). In this chapter, I describe the identification of multiple novel lncRNAs that are differentially expressed across the cervical-thoracic axial boundary in the developing chick embryo. I focus on one lncRNA in particular whose location in the genome is closest to that of the HOXC cluster, which has previously been identified as vital for thoracic development.

5.1.1. Classification of lncRNAs

Beginning in 1992, the total number of publications identifying lncRNAs steadily increased, with a sharp rise after 2011 (Quek *et al.*, 2015). Despite this, very few lncRNAs have been assigned functions (Quek *et al.*, 2015). Due to the lack of understanding of lncRNAs, they are still arbitrarily defined as RNA transcripts that are >200nts in length with no discernable protein-coding features. These transcripts are often polyadenylated with a 5' terminal methylguanosine cap, transcribed by RNA pol II and can have splice variations and histone modifications at their promoters similar to those found at mRNAs promoters (Gaiti *et al.*, 2015). Whilst only a fraction of lncRNAs have been characterized, those that have been explored have been linked to a variety of functions and are often differentially expressed throughout different stages of development, specific tissues or numerous disease states in a diverse range of species from plants to mammals (Quinn and Chang, 2016).

Established classes of RNAs have been assigned names based on their specific function. The large and well-established classes of RNAs, such as mRNAs, transfer (tRNAs) and ribosomal (rRNAs) all have particular roles in the production of proteins (Morris and Mattick, 2014). Some well-known classes of small RNAs include:

- micro (miRNAs) – 20-25nts, formed from single-stranded RNA that fold into hairpin structures to prevent translation or degrade mRNA transcripts through interactions with the Argonaute (AGO) protein component of an RNA-induced silencing protein complex (RISC) (Grosshans and Filipowicz, 2008; Morris and Mattick, 2014)
- small interfering (siRNAs) – 20-25nts, formed from double-stranded RNA and recognized for their roles in silencing transposons and viral infections via AGO-RISC interactions (Grosshans and Filipowicz, 2008; Morris and Mattick, 2014)
- piwi-interacting (piRNAs) – 25-30nts, formed from single-stranded RNA and associate with a subclade of AGO proteins termed PIWI proteins to carry out silencing of transposons in germ cells (Grosshans and Filipowicz, 2008; Morris and Mattick, 2014)
- small nucleolar (snoRNAs) – 60-300nts, formed from mRNA introns and act by guiding ribonucleoprotein complexes to target RNAs to induce chemical modifications in a site-specific manner (Dieci *et al.*, 2009; Falaleeva and Stamm, 2013)
- small nuclear (snRNAs) – 100-300nts, localized in eukaryotic cell nucleus and involved in RNA splicing (Morris and Mattick, 2014)

These have all demonstrated roles in gene regulation via specific interactions and are classified based on how they act (Morris and Mattick, 2014). The remaining RNA transcripts that do not fit well into any of these classifications have generally been assigned the term lncRNAs with a limit of being at least 200nts in length to separate them from the small ncRNAs. They have also normally not been functionally characterized. lncRNA transcripts have now been identified in diverse species, such as animals, plants, yeast, prokaryotes and viruses. Generally, they exhibit very limited sequence conservation compared to the other classes of RNA molecules (Ma *et al.*, 2013). This in part led to the original assumption that lncRNAs lack any functional biological relevance, as sequence conservation is typically linked to significant and usually conserved function (Ohno, 1972; Struhl, 2007).

The few lncRNAs that have been functionally characterized are associated with a multitude of biological processes, such as transcriptional activation, silencing, splicing, protein complex organization, cell cycle progression, and much more (Quinn and Chang, 2016). lncRNAs play crucial roles in transcriptional gene regulation with disruption to their function often leading to severe biological disorders. However, due to lack of information based on functional

characteristics, an early effort to classify lncRNAs attempted to group them based on genomic location relative to protein-coding genes. This divided lncRNAs into intergenic and intragenic, and then antisense, sense, intronic, and divergent relative to the nearby or overlapping protein-coding gene. Although some of these terms are still used, particularly intergenic, this normally conveys almost no information regarding function. Thus, this nomenclature results in lncRNAs with similar functions being given different, and therefore misleading, designations (Ma *et al.*, 2013).

Recent efforts have led to new categories based on length, physical association with protein-coding genes, association with other functional DNA elements, resemblance to mRNAs, association with repeats, association with biochemical pathways, stability, sequence and/or structure conservation, expression in different biological states, association with subcellular structures, or function (St Laurent *et al.*, 2015).

5.1.2. Identification of functional lncRNAs

A common problem with investigating the functions of lncRNAs is filtering the thousands of transcripts for those that are most likely to have biological functions from the inevitable subset that may represent transcriptional noise. This filtering is important due to the time, cost and effort that need to go into understanding how a single or group of lncRNAs may function. A variety of available technologies have been used to approach this issue. To begin with, transcription must be reliably detected and be independent of protein-coding genes. Transcripts that are within or overlapping protein-coding gene transcription, are often considered immature RNA and consolidated with the mRNA.

Once lncRNAs are annotated throughout a specific genome, transcripts most likely to have biological functions must be identified. Many lncRNAs have been called non-functional and by-products due to transcriptional noise from RNA pol II binding to weak promoters, experimental artifacts from possible contamination of residual genomic DNA, immature, unspliced introns or annealing of oligo-dT primers to adenine rich regions of other RNAs, rather than poly(A) tails (Louro *et al.*, 2009). This is because lncRNAs are frequently found to have very low levels of transcription and are shorter and less complicated with no splicing events or fewer exons/introns than mRNAs (Ravasi *et al.*, 2006). They are thought to arise from cryptic promoters within protein-coding gene regions, or intergenic regions that are devoid of histones, thus allowing access to transcription machinery that initiates at sequences that resemble promoters by chance (Struhl, 2007). Many lncRNAs also appear to be rapidly degraded (Baker and Parker, 2004) and may require more energy to prevent their transcription than to degrade.

Investigating if levels of transcript expression are altered in response to different conditions, tissue type or diseases is a possible method of investigating functions, although bearing in mind that the lncRNAs transcription could be altered due to chromatin environment also. One approach would be to identify the regulatory factors that are responsible for particular lncRNAs expression and to disturb their functions, leading to changes in expression levels, domains of expression, tissue type or time period of transcription of the lncRNA. The evolutionary preservation of the lncRNA expression could then be investigated in closely related species. It would be expected that if their expression is not under selective pressures then there would be no need to maintain these patterns (Chodroff *et al.*, 2010; Hezroni *et al.*, 2015). Investigations into tissue specificity have found that lncRNAs exhibit a much higher tissue specificity, regardless of levels of expression, than protein-coding genes, leading many to believe that they are indeed functionally relevant (Cabili *et al.*, 2011).

Evolutionary sequence conservation is considered a reliable indicator of conserved functional biological roles for protein-coding genes, as normally they need to preserve an open reading frame and much of their amino acid sequence. Further debate for lack of evidence for the majority of lncRNAs possessing function is that they generally exhibit low evolutionary sequence conservation, when compared to protein-coding genes, although overall, lncRNAs have been found to have more conservation than introns or random intergenic sequences (Guttman *et al.*, 2009). However, it does not necessarily follow that lncRNAs continue this same pattern of nucleotide conservation. Rapid turnover of sequence, along with a continuous process of generation and loss of novel and existing lncRNAs, may better resemble the subtler rate of phenotypic variation between closely related species. Changes in the non-coding, regulatory regions could contribute to the small and enormous differences in body morphologies found throughout the metazoan kingdom.

Possibly the main consideration during the identification of bona fide lncRNAs is to confidently dismiss the possibility that they encode a functional protein. This is an ongoing debate between researchers as the majority of tools used to detect the potential of an RNA transcript to produce a protein are based on bioinformatic predictions built on current knowledge of the central dogma of biology. These tools assess an RNA sequence based foremost on possession of a continuous stretch of sequence that begins with an initiator codon for methionine and that stretches across some distance to a stop codon, known as an open reading frame (ORF). By chance, short ORFs will occur in any 1000nt stretch of DNA, the average length of lncRNAs. However, based on the majority of known proteins having been found to contain ORFs of

>300nts, this was used as the conventional ORF length to distinguish mRNAs from ncRNAs (Dinger *et al.*, 2008). It was later realized that this was largely dependent on length of sequence and that very long lncRNAs would be classed as mRNAs even if their lncRNA functions were well established (Prasanth and Spector, 2007). For example, Xist contains an ORF of just under 900nts and was classified in databases as a protein-coding gene for 15 years before being recognized as a lncRNA (Brockdorff *et al.*, 1992). Modern algorithms now calculate minimum ORF cut-offs in a length dependent manner for a more reliable assessment of the protein-coding potential, although precaution is still required as there have been examples of misclassification. For example, the tarsalless (*tal*) gene was initially classed as ncRNA as it contained very small ORFs but was later found to be translated into several 11aa peptides (Galindo *et al.*, 2007).

5.1.3. Methods for investigating lncRNA functions

Once transcripts have been bioinformatically assessed and do not appear to fall into another RNA classification, leaving them in the category of lncRNAs, there is yet to be established a 'gold standard' laboratory method for further investigations. Some traditional approaches, such as loss-of-function and gain-of-function experiments, can be used to experimentally investigate functionality of lncRNAs. To investigate the effects of loss-of-function, a lncRNA can be inhibited with small-interfering RNAs, antisense oligonucleotides (ASO), locked ASOs or morpholinos (Marin-Bejar, 2015). However, first the sub-cellular localization of the lncRNA would need to be established as RNAi machinery is located in the cytoplasm. Therefore, they been found to be more effective at targeting lncRNAs found in the same cellular compartment, such as OIP5-AS1 and DANCER (Lennox and Behlke, 2016). ASOs were more successful at knocking down lncRNAs localized to the nucleus, such as MALAT1 and NEAT1 and for those found in both compartments the ASOs were found to be more effective (Lennox and Behlke, 2016). However, differences in strategies used to assess function mean multiple gene manipulations should be employed to obtain a reliable assessment of function. For example, RNAi knockdown of Evf-2 led to the conclusion that it was necessary for activation of Dlx5/6 (Feng *et al.*, 2006), whereas transcriptional termination had the opposite effect on the same gene (Bond *et al.*, 2009). However, this could be due to the RNAi being carried out in cell culture and the termination experiments *in vivo* (mice). There was similar confusion over the function of lincRNA-p21 when RNAi demonstrated that it was recruiting protein complexes to chromatin in trans (Huarte *et al.*, 2010), but ASO and promoter deletion instead showed a *cis* mechanism of regulation of the adjacent gene, p21 (Dimitrova *et al.*, 2014). Other efficient methods involve targeting specific loci with nucleases, such as zinc-finger nucleases (ZFNs) transcription activator-like effector nucleases (TALENs) or clustered regulatory interspaced short palindromic repeats associated endonuclease (CRISPRCas9) to either partially delete a region of the lncRNA, or the whole

section spanning the gene or just the promoter (Cheng *et al.*, 2013). This can be further adapted to insert sequences at cut sites that may disrupt transcription (Gilles and Averof, 2014).

Ectopic expression of lncRNAs may mimic the endogenous functions of the transcript allowing measurements that may indicate if there are genes that the lncRNA is acting upon and possibly if it is a negative or positive regulation. These results often require verification using other methodologies, as theoretically expressing a gene outside its endogenous spatiotemporal restrictions could allow interactions with molecules that are not otherwise available. Visualization techniques, such as nascent transcript fluorescent *in situ* hybridization (ntFISH) can be used to provide crucial information about tissue specificity, colocalization with other transcripts or proteins, if the lncRNA localizes to the nucleus or cytoplasm and if it is transcribed from one or both alleles. For example, fluorescent labelling of Xist showed that it was transcribed from and localized at the inactive X chromosome, thus providing an insight into its function (Brown *et al.*, 1992; Clemson *et al.*, 1996). Visualization of NEAT1 demonstrated that it was highly abundant in paraspeckles (Clemson *et al.*, 2009) and when both NEAT1 and MALAT1 were visualized, they were found to be associated with SC35 nuclear speckles, leading to the finding that they were involved in mRNA metabolism (Hutchinson *et al.*, 2007).

More information comes from analysis of interactions between lncRNAs and other molecules. lncRNAs have been found to interact with DNA, RNA and proteins, and identification of these factors with a lncRNA can establish its likely mechanisms. For example, a lncRNA from the DHFR loci can form a triplex with the DNA at the DHFR promoter and prevent Pol II transcription (Martianov *et al.*, 2007) and lncRNAs have been found interacting with mRNAs, such as half-STAU1-binding site RNAs ($\frac{1}{2}$ -sbsRNAs) with 3'UTRs of two mRNAs (Gong and Maquat, 2011) and TINCR with several mRNAs (Kretz *et al.*, 2013). A wide variety of protocols have been developed in recent years to acquire knowledge of factors that are bound by lncRNAs utilizing variations of cross-linking and immunoprecipitation. In some methods a protein is used as bait to capture RNA bound by the protein (Darnell, 2012), which assumes knowledge of the lncRNAs protein binding partners. In other techniques antisense lncRNA is used to capture the sense RNA molecule along with the DNA, RNA and/or protein (Chu *et al.*, 2012; Engreitz *et al.*, 2015). However, this requires large amounts of material and may not capture the lncRNA in its native folded structure that allows it to bind to its interacting partners. If the DNA is used as bait in chromatin conformation capture technologies (Dekker *et al.*, 2013) then this method needs to be combined with other methods to identify proteins binding to the lncRNA and therefore alone does not provide a lot of mechanistic information. Furthermore, many of these techniques have been developed for specific analysis and overall designed to analyse large amounts of material

that can be generated from the use of cell lines. However, *in vivo* analysis is more difficult as the expression levels of lncRNA are typically quite low and specific to small subsets of cells within an organism (Yang *et al.*, 2015).

5.1.2. Identified functions of lncRNAs

Several lncRNAs are known to be fundamental throughout development, with the ability to influence genes necessary for dosage compensation, cell differentiation, organogenesis and body pattern specification (Fatica and Bozzoni, 2014). Many of the genes that are involved in developmental gene regulation have also been associated with human diseases such as breast, skin, liver, colon and prostate cancers, genetic disorders, diabetes, neurodegeneration and neurological disorders (Di Gesualdo *et al.*, 2014). The diseases are thought to be linked to developmental processes by the dysregulation of lncRNAs that regulate cell proliferation or apoptosis during development, leading to inappropriate expression of genes that control these events.

Functional investigations of a lncRNA requires evidence to determine if the transcript itself is functional, as transcription of lncRNA loci can be all that is required to mediate gene regulation by regulating functions associated with the underlying DNA. There are now several lncRNAs for which the act of transcription itself seems able to regulate activity of gene in *cis* via interference with TFs, nucleosome repositioning or affecting promoter associated histone modifications (Kornienko *et al.*, 2013). There are still very few examples of this method of lncRNA regulation. One example uses transcriptional interference via displacement of RNA pol II machinery, reported for the very long mammalian lncRNA, Airn, found to overlap the Igf2r gene promoter (Latos *et al.*, 2012). Airn is thought to utilize multiple unknown silencing mechanisms to mediate silencing of paternal alleles of Igf2r, Slc22a3 and Slc22a2, at different developmental stages. However, the Igf2r promoter and not the Slc22a3 or Slc22a2 promoters are overlapped by Airn transcription, and it was found that alterations could be made to the transcript length that would therefore eliminate the lncRNA products function, but it would maintain silencing as long as it was transcribed through the Igf2r promoter. Similar examples of this mechanism have also been observed in two instances in the Hox complex in *D. melanogaster*. In one case researchers reasoned that transcriptional elongation of the lncRNA, bithoraxoid (bxd) transcripts were likely to repress the neighbouring gene, Ubx, as they were not transcribed in the same cells of the developing embryo and deletions of bxd led to ectopic expression of Ubx in regions usually occupied by bxd's expression (Petruk, 2006). Further upstream in the Hox complex, another very long lncRNA, infraabdominal-8 (iab-8), is thought to interfere with the promoter of abdominal-

A (*abd-A*), to repress expression in combination with the repressive effects from a miRNA produced from within *iab-8*, *mir-iab-4* (Gummalla *et al.*, 2012).

Instead of interfering directly with TFs at gene promoters, lncRNA transcription can also affect gene expression by changing the nucleosome density in promoter or enhancer regions, thus facilitating or restricting TF access to DNA. This has been shown to be the case for the transcription of the *S. cerevisiae* lncRNA, *SRG1*, which represses the adjacent gene *SER3*. The silencing is a result of transcription across the *SER3* promoter that increased nucleosome occupancy at the DNA, whilst no function could be identified for the lncRNA product (Hainer *et al.*, 2011). The nucleosomes are deposited behind RNA pol II as it transcribes through *SRG1* in a rate independent manner that requires the presence of an elongation factor Spt2 for the recycling of old histones to reform a repressive nucleosomal structure (Thebault *et al.*, 2011). Although this mechanism of gene regulation by nucleosome repositioning has only been supported in limited studies and for quite long lncRNAs, it does reflect the general convention of lncRNA remodelling chromatin that has gained the most support for lncRNA function.

Currently, the most commonly observed function of lncRNAs in gene regulation is via their interaction with chromatin modifying complexes. A common paradigm that has underpinned much of what we now know of lncRNA function is an association with a highly evolutionary conserved repressive complex, PRC2, a histone methyltransferase. This was first demonstrated in *cis* as the means by which *Xist* establishes inactivation on the X chromosome of mammalian females to mediate dosage compensation between males and females (Borsani *et al.*, 1991; Brockdorff *et al.*, 1991; Brown *et al.*, 1991).

lncRNAs actions in regulating dosage compensation are highly conserved, for example in organisms as diverse as fruit flies and marsupials. The marsupial, *Monodelphis domestica*, also has one X chromosome silenced in order to maintain dosage and match transcript levels of the single copy of the X chromosome inherited by males. The lncRNA required to do this is called RNA-on-the-silent-X (*Rsx*), a large 27 kb transcript that is expressed only in female *M. domestica*. It appears to function in largely the same way as *Xist*, in that it is transcribed from and coats only the inactive, paternal X chromosome. In *D. melanogaster* dosage compensation is reliant on the up regulation of gene expression from the single X chromosome in males to match the female expression levels from two active X chromosomes. This is now understood to be coordinated by a male-specific lethal complex (MSL) and male-specific RNAs on the X, *roX1* and *roX2*. These lncRNAs are transcribed from the male X chromosome and co-transcriptionally incorporated into the MSL (Meller *et al.*, 2000; Meller *et al.*, 1997) in a rate dependent manner (Kelley *et al.*,

2008). This demonstrates that chromatin reorganization via interactions of lncRNAs are remarkably conserved in exceptionally evolutionary divergent organisms and are likely to be similar throughout the animal kingdom.

The direct interaction of lncRNAs with PRC2 has been commonly observed, with another established multifaceted lncRNA complex found in the plant species *A. thaliana*. This complex is responsible for regulating vernalization, a process that regulates flowering in Spring response to prolonged exposure to cold at the Flowering locus C (FLC) gene (Swiezewski *et al.*, 2009). FLC is a TF that represses the transition to flowering and is epigenetically (increased H3K27me3 and decreased H3K36me3) silenced by a complex of PRC2 and plant Homeodomain proteins (PHD) in cold periods (Kim *et al.*, 2009). The chromatin at FLC is switched to a permissive chromatin state by trithorax homologs ATX1 and ATX2 depositing H3K4me3 and H3K4me2 respectively allowing transcription of FLC (Pien *et al.*, 2008). Two lncRNA transcripts are transcribed from the FLC loci, COOLAIR and COLDAIR. COOLAIR originates in the antisense orientation, fully covering the FLC locus, to produce 2 non-coding isoforms. These isoforms are termed ASI and ASII and are produced through splicing and alternative poly(A) signals. Interestingly, COOLAIR is regulated by differential stabilization from a homeodomain protein, NDX1 via formation of an Rloop, a 3-stranded structure of RNA; DNA and single-stranded DNA (Sun *et al.*, 2013) that have also been shown to form heterochromatin in mammals. Increased expression of ASI lead to increased FLD (homolog of Lysine Specific Demethylase, LSD1) mediated demethylation of H3K4me2 that in turn represses transcription of FLC and the AS II form is still not well understood.

Another lncRNA is transcribed from within the first intron of FLC in the sense orientation, COLDAIR (Heo and Sung, 2011). This transcript is expression later than COOLAIR and is not polyadenylated or alternatively spliced, but does have a 5' cap. COLDAIR, unlike COOLAIR, is responsible for physically interacting with CURLY LEAF (an Enhancer of Zeste homolog), a PRC2 component, and is hypothesized to form a scaffold that binds the complex and recruits it to FLC to establish silencing by H3K27me3. The COLDAIR and COOLAIR nomenclature is derived from another lncRNA, Hox transcript antisense (HOTAIR), a 2.2 kb lncRNA identified in humans that also interacts with PRC2 to trimethylate H3K27 (Rinn *et al.*, 2007). HOTAIR was the first lncRNA shown to regulate genes in trans, as it is transcribed antisense to HOXC11, a component of the HOXC cluster on chromosome 12 but leads to silencing of a cluster of HOXD genes on chromosome 2, specifically HOXD8, 9, 10 and 11, without affecting the HOXC genes in the region it is transcribed from (Khalil *et al.*, 2009). Along with guiding PRC2 to target gene promoters for silencing, HOTAIR also uses another distinct binding domain to interact with LSD1, a histone

demethylase that forms part of the CoRepressor for element-1-silencing transcription factor CoREST complex, specifically removes methylation marks from H3K4 to further contribute to gene silencing. The HOTAIR transcript was shown to act as a scaffold, binding PRC2 at the 5' end and LSD1 at the 3' end to recruit them to its specific gene targets for epigenetic silencing (Fig.5.1) (Tsai *et al.*, 2010). These interactions with PRC2 and LSD1 are mirrored by COOLAIR in plants, suggesting this is a highly conserved, ancient mechanism of gene regulation that is present in both plant and animal kingdoms.

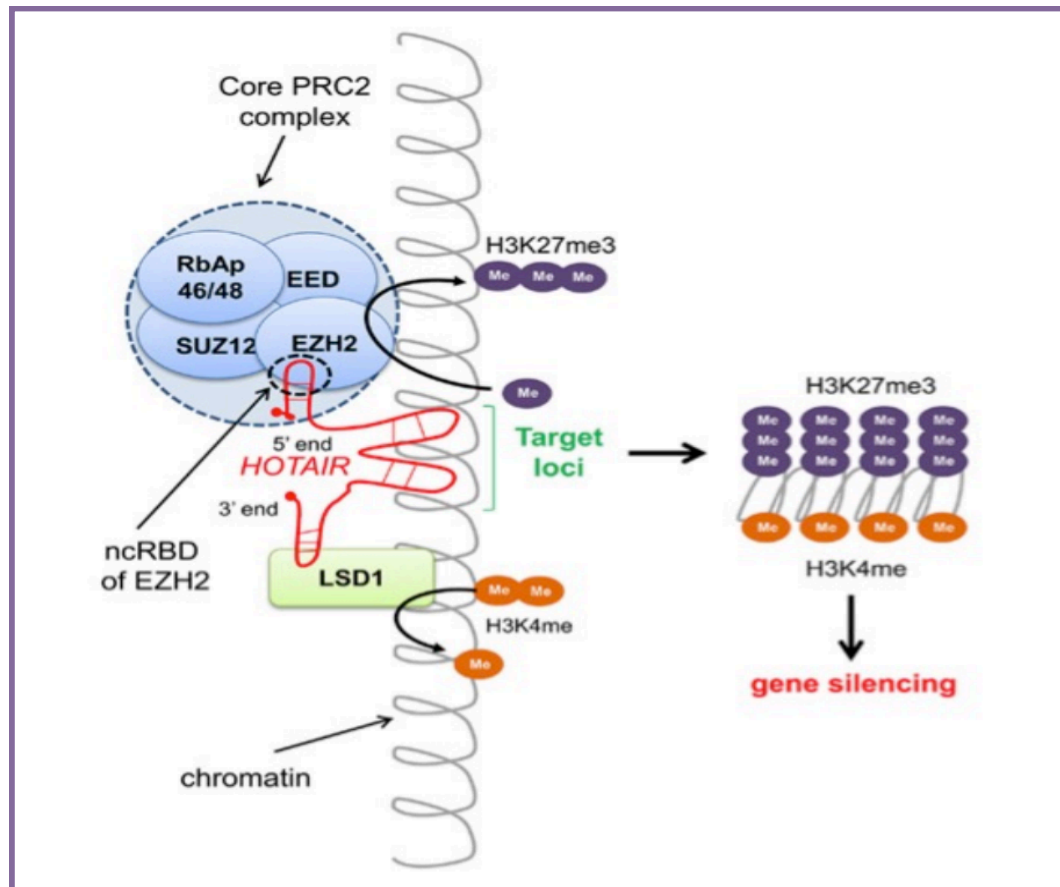


Figure 5.1: Interactions of lncRNA HOTAIR to induce silencing to target loci. HOTAIR interacts with LSD1 at the 3' end and the EZH2 component of the PRC2 complex at the 5' end leading to H3K27me3 and H3K4me modifications to induce silencing of target loci on a different chromosome than the HOTAIR locus, thus acting in trans. ncRBD = non-coding RNA binding domain. Image from (Marsh *et al.*, 2014).

Other functional lncRNAs have also been identified in mammalian HOX loci, including HOXA transcript at the distal tip (HOTTIP), HOXA11 antisense RNA (HOXA11-AS), HOXA3- AS, HOXA6- AS, HOXA transcript antisense RNA, myeloid-specific 1 (HOTAIRM1) and FRIGIDAIR (Quek *et al.*, 2015).

5.2. Results and Discussion

5.2.1. Identification of lncRNAs from RNAseq dataset

Using RNA sequencing (as described in Chapter 3), I identified over 400 novel lncRNAs that were differentially expressed across the cervical-thoracic axial boundary in chick embryos, 18 of which were included in the list of 343 differential genes identified in Chapter 3. More specifically, of the top 25 sequences out of 343 identified as differential, 10 of these were novel lncRNAs (Table 5.1). This highlights a previously unknown and possibly very significant contribution of lncRNAs to the fixation of cell fate to axial identity.

Table 5.1: List of top 10 differential novel lncRNAs identified from RNAseq dataset.

No.	lncRNA Ensembl ID	UCSC coordinates	Forward Primer	Reverse Primer
1	ENSGALG00000054525	chr6:26,598,442-26,600,995	ACTTCTCCTCTCACATTGCC	AGAGGGAGCTTGCCTAG
2	ENSGALG00000050591	chr3:41,647,509-41,688,626	ACCACTCACCAGCCAAAACA	TTGTGATGTGAAGCCTCGCT
3	ENSGALG00000053345	chr33:7,682,561-7,686,502	CCTTTCTCCTCCCAGGGTTC	AATTAGCCCCGGGTGAGG
4	ENSGALG00000051998	chr1:27,099,983-27,107,558	ATTTCAGAGAGCGTGGCAA	AGCTCAACCAGAGTGATGCC
5	ENSGALG00000047842	chr4:76,888,441-76,901,092	GCCCTTCTAATATGGCCGGT	AGGGAACCTTGCTCCCTATTG
6	ENSGALG00000054479	chr4:31,166,169-31,179,406	CGTTGACTCTCAAAGCCCCT	CTCCAGGAGTGCTCCTACT
7	ENSGALG00000050453	chr4:81,825,853-81,843,984	CCACTTCTGCCTGGACGATT	ACTGTTGACCGCATCCCATT
8	ENSGALG00000052325	chr2:21,442,791-21,459,430	ACCCAAGGCATCACTGTAGC	ACAGGATAGGACAGTCCCCC
9	ENSGALG00000051195	chr14:8,159,547-8,166,204	ACATTGAGTCAGGCACAGCA	CATTCCAGCCCCATGCACTA
10	ENSGALG00000054927	chr5:45,330,479-45,333,047	TGGTGGGTCCATTCTTCTGC	TCTCCATGTGCTACTCGCAC

All sequences have a PolyA tail, which is how they were selected in the RNA sequencing protocol (as described in Chapter 2). None of the sequences, despite their considerable length (ranging from 2.5kb to 40kb) overlap any protein-coding genes within the genome. Furthermore, analysis using online resources, such as predictprotein.com (Bernhofer *et al.*, 2021), was unable to predict protein structure or function for any of the 10 sequences. This made it very difficult to identify any of their function as I could not speculate what genes they may interact with.

5.2.2. Visualising lncRNAs in the chick embryo

To date, biomarker efforts have mainly focused on protein-coding genes, which comprise only a subset of all transcribed genes (Birney *et al.*, 2007; Kapranov *et al.*, 2007). Among the more than 90% of transcription that generates noncoding RNAs, long noncoding RNAs (lncRNAs) most closely resemble protein-coding genes in that they are transcribed by RNA polymerase II, polyadenylated, and associated with specific epigenetic signatures (Guttman *et al.*, 2009). Based on my findings from the RNAseq dataset, I sought to make the first step in characterising these novel lncRNAs by determining their expression patterns. Using ISH in HH16 chick embryos would thus determine their association with the transition from cervical to thoracic identity. The results for 9 out of 10 novel lncRNAs examined are displayed in Figure 5.2. Unfortunately, I was unable to perform ISH for the final novel lncRNA on the list, lncRNA no. 10, ENSGALG00000054927 (Table 5.1). This was due to the original primer pair designed failing to produce a working ISH probe and time constraints.

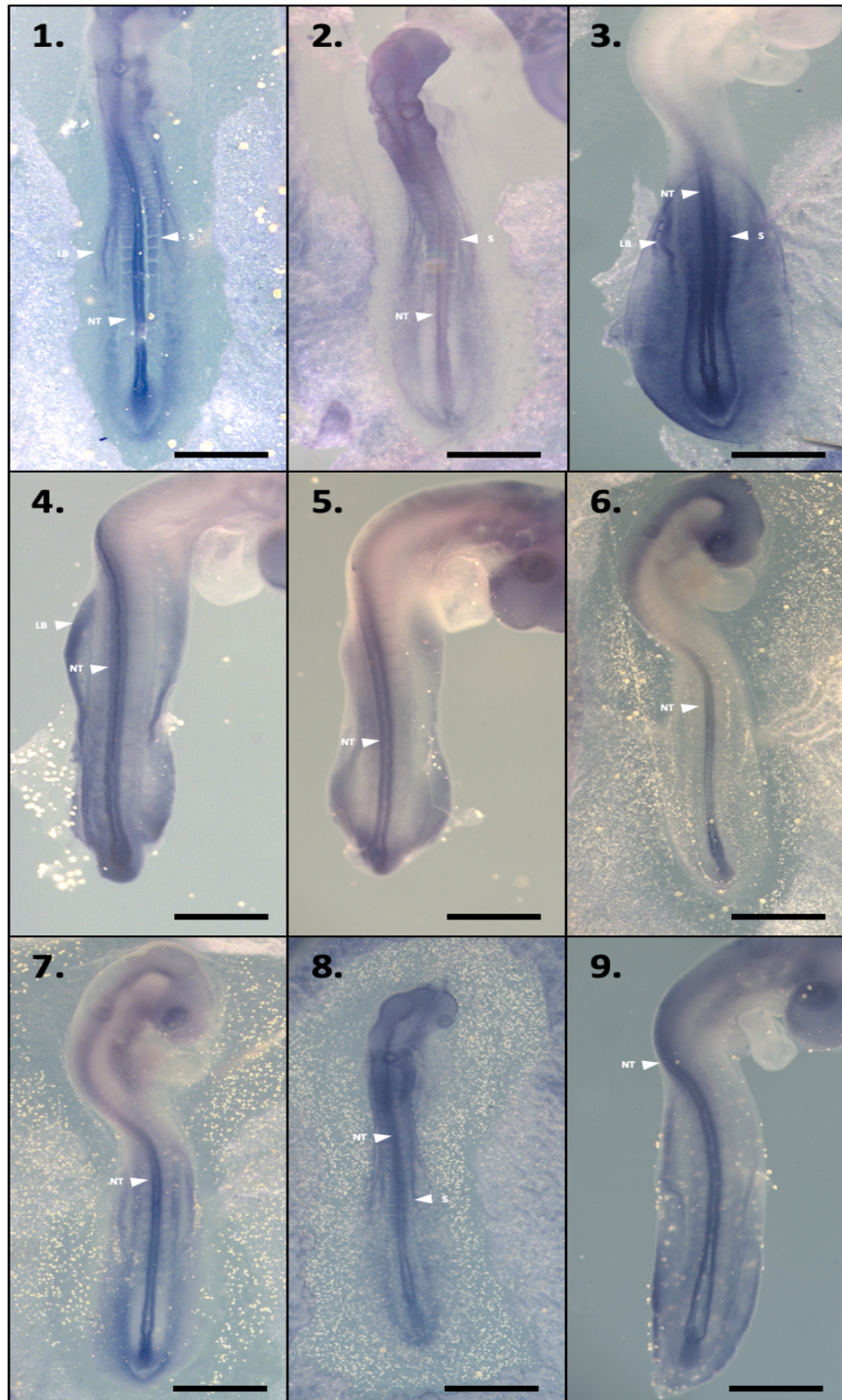


Figure 5.2: Whole mount in situ hybridisation of nine novel lncRNAs identified from the RNAseq dataset. Embryo stages ranging from HH14-18. The numbers correspond to Table 5.1, which also shows the Ensemble I.D. for each lncRNA. White arrows indicate important areas of expression: neural tube (NT), somites (S), and limb bud (LB). Size bar = 0.5mm.

All nine of the lncRNAs tested showed restricted expression patterns within the chick embryos. Four out of the nine lncRNAs tested displayed expression predominantly in the somites (Figure 5.2.1;2;3;8), with two lncRNAs displaying clear differential expression in the somites across the C-T boundary (Figure 5.2.1 & Figure 5.2.3). Figure 5.2.2 also shows a very slight increase in expression in the more posterior thoracic somites, whereas Figure 5.2.8 does not appear to be differentially expressed across the somites. The remaining lncRNAs are lacking any expression in somitic tissue, instead primarily being expressed along the neural tube and in the limb buds, except for Figure 5.2.6, which only shows expression in the neural tube. Due to the short fragment lengths used when generating the ISH probe, the complete expression pattern for some lncRNAs may be lost due to splicing as well as the presence of difference isoforms, thus possibly explaining the lack of somitic signal for these embryos.

5.2.3. A closer inspection of novel lncRNA ENSGALG00000053345

Of the 10 lncRNAs selected for further characterisation, only one sequence, ENSGALG00000053345, is located on the same chromosome as the HOXC cluster of interest for this project, ~16Kb downstream of *HOXC4* (Figure 5.3). Further inspection of novel lncRNA ENSGALG00000053345 within the ATACseq dataset on the UCSC genome browser also identified a differential peak within what appears to be the promoter region, corresponding with the differential pattern identified in the RNAseq dataset (low in cervical sample and high in thoracic samples).

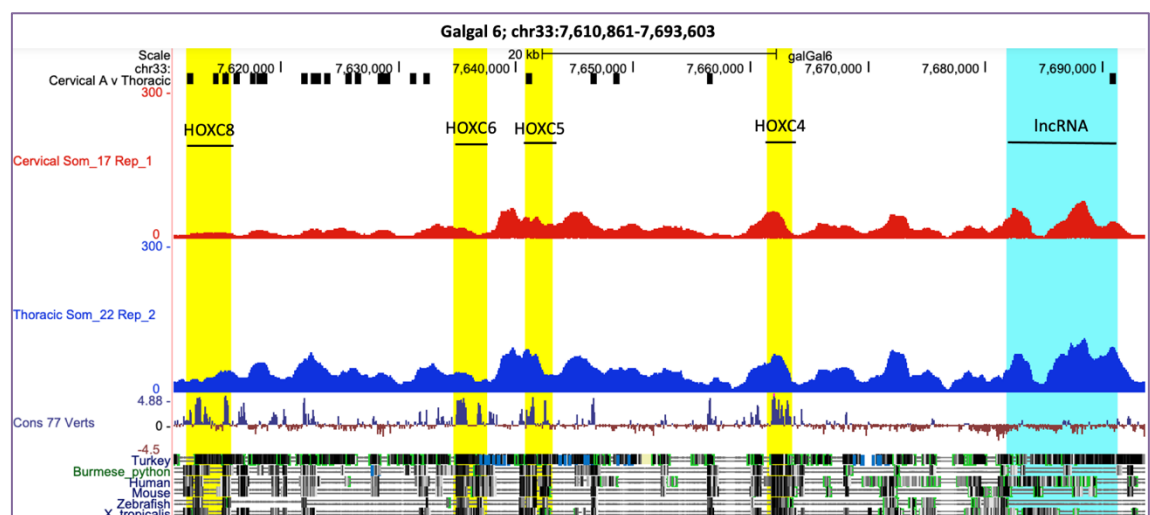


Figure 5.3: Genomic location of novel lncRNA ENSGALG00000053345 in relation to the HOXC cluster. ATAC-seq profile at the *HOXC* cluster. Yellow boxes indicate genes across the cluster with black bars indicating the specific gene name. Blue box indicates the novel lncRNA identified, ENSGALG00000053345. Black bars indicate location for all differential peaks identified across the genome section imaged. Conservation across selected vertebrate species is also indicated by the black columns at the bottom of the image.

Alongside both the similar log fold change to HOXC6 in the RNAseq dataset and the neighbouring location in the genome, the ISH expression pattern for this novel lncRNA is identical to the expression pattern to HOXC6 (Figure 5.4). Both HOXC6 and this lncRNA are first expressed in somite number 19/20, which is the exact location where cervical axial identity halts and thoracic identity begins. As well as this, both show a more anterior expression in the neural tube, with expression in the neural tube corresponding to a shift of about 2 or 3 somites than the somitic expression pattern. Expression is also seen throughout the presomitic mesoderm and along the apical ectodermal ridge of the limb bud. Due to this, I hypothesis that this lncRNA has a similar, if not overlapping, function as HOXC6 and thus has a critical role in the determination of the thoracic domain in axial development.

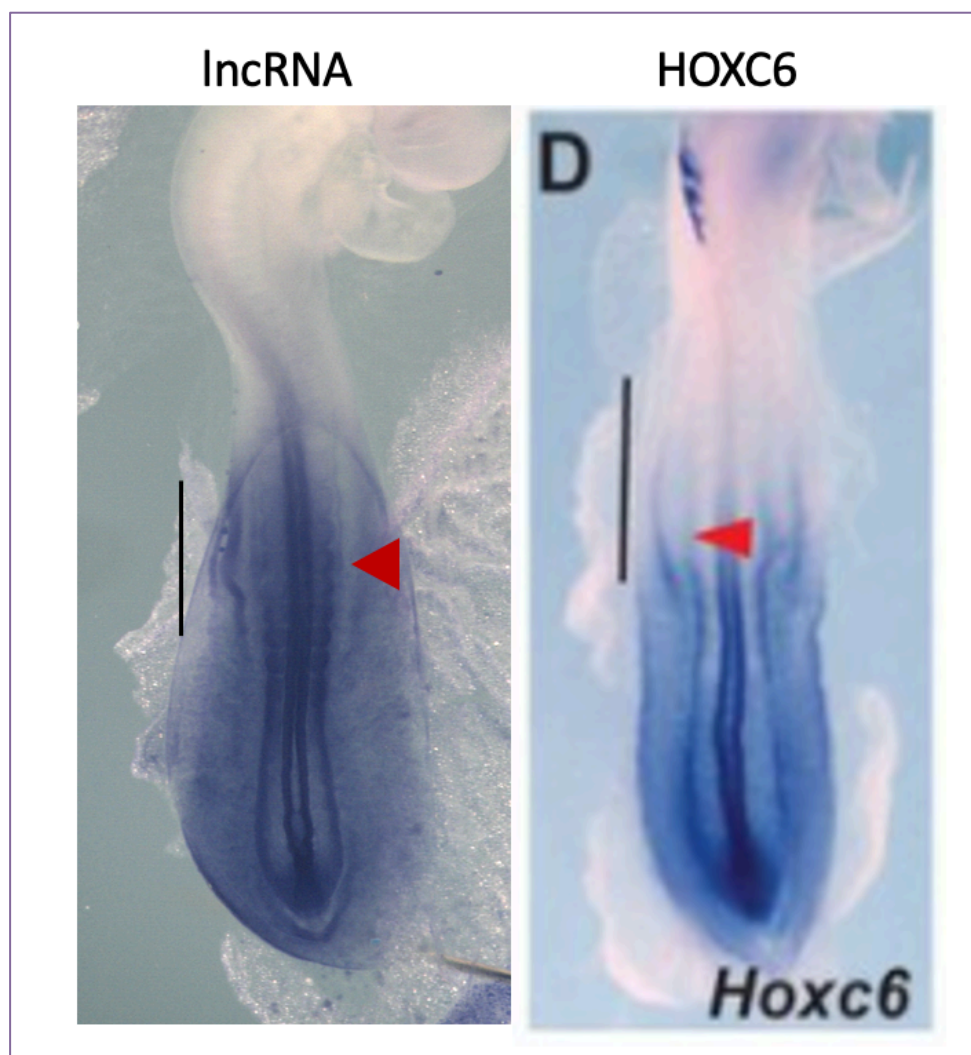


Figure 5.4: Comparison of ISH between HOXC6 and novel lncRNA. The red arrow indicates the level where expression begins in the somites, also marking the boundary between cervical and thoracic somites. The black bar represents the length of the forelimb bud.

A possible function for this lncRNA is that it may act as a microRNA sponge. A search for miRNA seeds within the sequence identified multiple miRNAs that had more than 60 possible seed sites within the sequence (Table 5.2). Many of these miRNAs are known to be involved in vertebrate evolution (gga-miR-190a-5p, gga-miR-190b-5p, gga-miR-19b-3p, and gga-miR-19a-3p) (ICGSC, 2004) and gastrulation (gga-miR-6547-5p, gga-miR-6700-5p, and gga-miR-6708-5p) (Shao *et al.*, 2012), suggesting this lncRNA could act as a sponge to modulate the activity of these microRNAs.

Table 5.2: List of miRNA targets within novel lncRNA ENSGALG00000053345.

miRNA ID	No. of possible seed sites
gga-miR-6547-5p	82
gga-miR-6700-5p	82
gga-miR-190a-5p	71
gga-miR-3535	71
gga-miR-190b-5p	71
gga-miR-6708-5p	69
gga-miR-1639	67
gga-miR-19b-3p	65
gga-miR-19a-3p	65
gga-miR-12236-3p	65

5.2.4. Interaction of novel lncRNA ENSGALG00000053345 with *HOXC6*

To investigate further the possible function of this novel lncRNA, I wanted to determine if the two genes important for the change in axial identity across the cervical and thoracic boundary interacted with this lncRNA in any way. Four guides were designed to target the promoters of both *HOXC6* and *HOXC8* (Appendix 1) individually in order to increase their expression using a modified CAS9 carrying a strong transcriptional activator domain, VP16, in a technique called CRISPR-ON. CRISPR-ON is a powerful RNA-guided transcriptional activator system that can turn on specific gene expression (Giménez *et al.*, 2016).

To confirm the technique was successful for *HOXC6*, I initially electroporated wild type HH4 chick embryos with the guides targeting *HOXC6*. Following a 24hr incubation, whole mount *in situ*

hybridisation was carried out on the injected embryos using the *HOXC6* probe in order to identify if any ectopic expression had occurred. Figure 5.5 below shows the successful appearance of ectopic *HOXC6* expression in the injected HH10 embryos, particularly in extraembryonic regions. HH10 embryos normally would not have any *HOXC6* expression, as confirmed in the wild type embryos that were not injected with the guides. This suggested the CRISPR-ON approach was successful and the guide RNAs were working as expected. Embryos injected with the *HOXC6* CRISPR-ON guides at HH4 then underwent whole mount ISH using the probe detecting lncRNA ENSGALG00000053345 probe. This showed that expression of lncRNA ENSGALG00000053345 increased, showing expression in a much younger embryo where it normally would not be detected. Thus, ectopic expression of ENSGALG00000053345 correlated with ectopic expression of *HOXC6*.

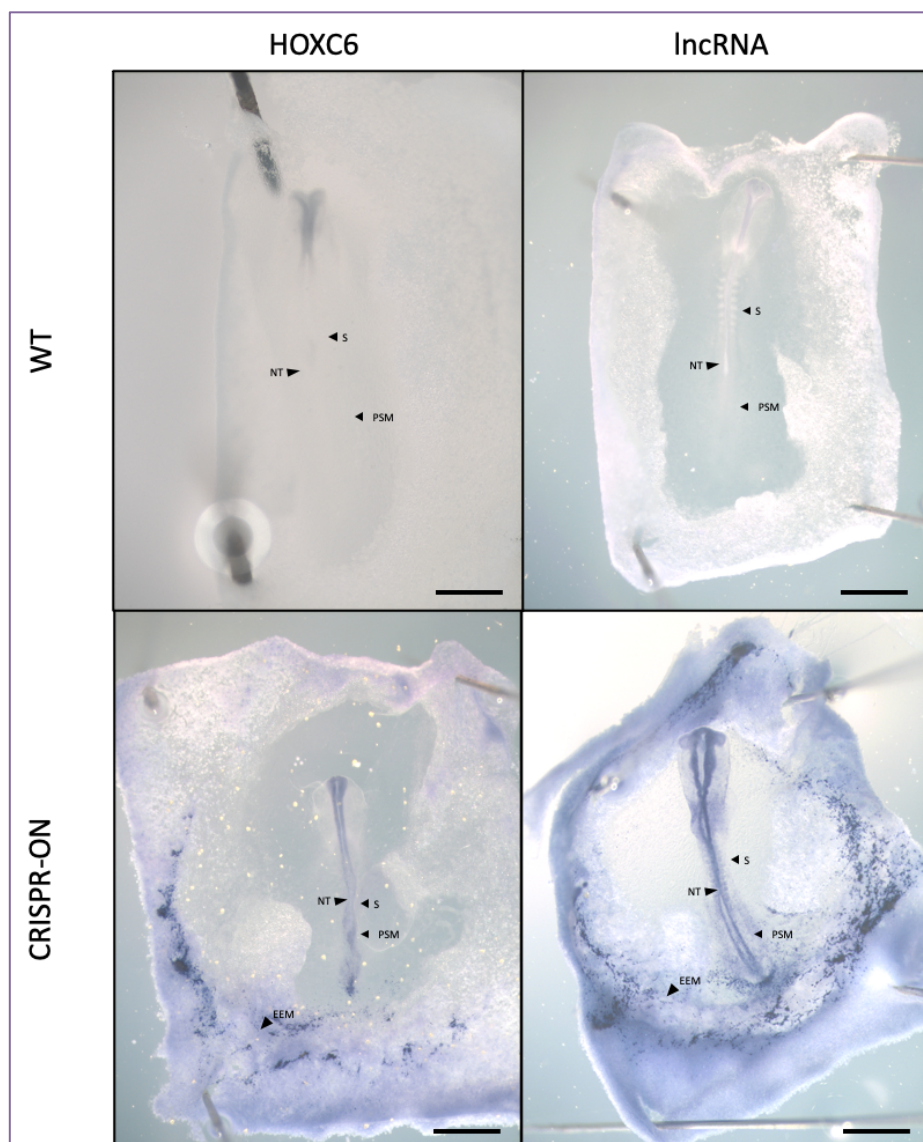


Figure 5.5: Visualisation of expression following CRISPR-ON of *HOXC6*. Black arrows highlighting different tissues where ectopic expression is observed: neural tube (NT), somites (S), presomitic mesoderm (PSM), and extraembryonic membrane (EEM). Scale bar = 0.5mm.

Upon confirming an increase in lncRNA ENSGALG00000053345 expression following *HOXC6* overexpression using CRISPR-ON, I wanted to confirm that this interaction was specific to *HOXC6* and not also an interaction that would occur with a gene of a similar expression pattern and function, such as *HOXC8*. CRISPR-ON was carried out once more, except this time it was *HOXC8* promoter that was targeted. Figure 5.6 shows that the experiment was also successful when using guides for *HOXC8*, with ectopic expression appearing in much younger wild type HH10 embryos where there would normally not be expressed. Following injection of the guides for *HOXC8* CRISPR-ON, whole mount ISH for the novel lncRNA ENSGALG00000053345 was carried out and this time no ectopic expression was seen in the injected embryos. This suggests there is no interaction between *HOXC8* and lncRNA ENSGALG00000053345.

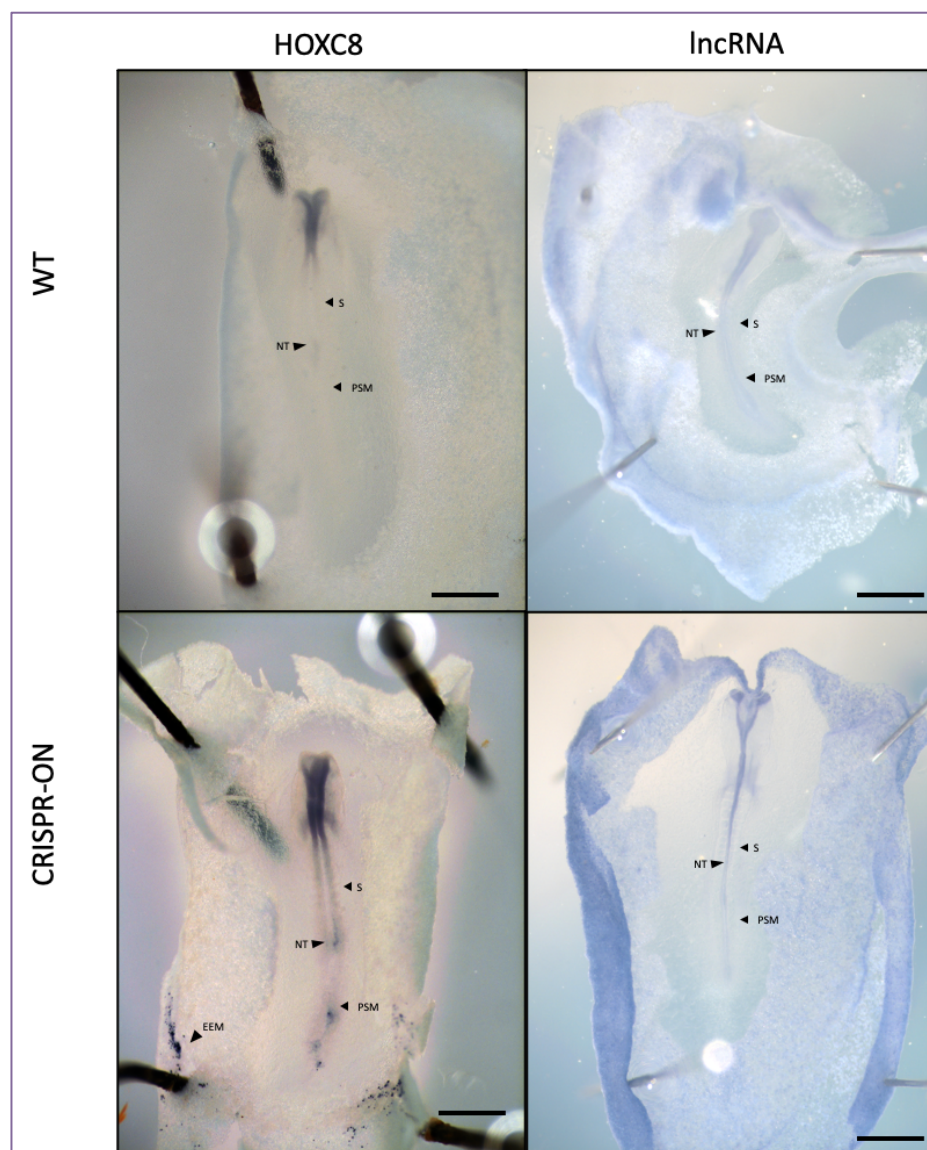


Figure 5.6: Visualisation of expression following CRISPR-ON of *HOXC8*. Black arrows highlighting different tissues where ectopic expression is observed: neural tube (NT), somites (S), presomitic mesoderm (PSM), and extraembryonic membrane (EEM). Scale bar = 0.5mm.

This experiment indicates that there is a very specific interaction between *HOXC6* and this novel lncRNA, as somites transition from a cervical axial identity to a thoracic axial identity. Specifically, *HOXC6*, which shares the same expression pattern to the lncRNA mediates this interaction and not *HOXC8*, which is also speculated to play a major role in the C-T transition. This shows that *HOXC6* must have a much greater role in this transition, specifically with regards to musculoskeletal development, than previously thought and should not be ignored, as past studies have mainly focused on *HOXC8* when investigating thoracic vertebrate development. Perhaps *HOXC8* is more crucial for central nervous system development than *HOXC6*, as this research focused on somites, not neural tube development.

5.2. Conclusion

In summary, multiple novel lncRNAs were identified as differentially expressed in the RNAseq dataset across the cervical-thoracic boundary in the chick embryo in this chapter. First, *in situ hybridisation* was successfully carried out on the top nine novel lncRNAs identified, with four of these showing significant somitic expression and two of those showing an absence of expression in cervical somites and onset of expression in thoracic somites.

One novel lncRNA, ENSGALG00000053345, was found located downstream of the *HOXC* cluster and showed a very similar expression pattern to *HOXC6*. This lncRNA was also identified as differential in the ATACseq dataset, with a differential peak identified within its promoter, which was more accessible in thoracic somites compared to cervical.

A search for microRNA seed sequences within this lncRNA sequence suggests it could act as a miRNA sponge, with the most frequent miRNAs with target sites within the lncRNA having a relation to both gastrulation and vertebrate evolution. Thus, modulation of several miRNAs could be one possible function of ENSGALG00000053345.

Further investigation into the function of lncRNA ENSGALG00000053345 included increasing the expression of *HOXC6* within younger embryos (HH8-10) using the CRISPR-ON technique. After increasing the expression of *HOXC6*, the expression of the novel lncRNA identified also increased, showing expression in a much younger embryo where it normally would not be expressed. This suggests that *HOXC6* interacts with this novel lncRNA. Future work is needed to further characterise this lncRNA as well as to identify the functions of the remaining lncRNAs identified in this chapter. For example, using ATAC-Seq to identify the promoter region for

ENSGALG00000053345, guides could be designed to carry out CRISPR-ON for this lncRNA and examination of the long-term effect on the development of the embryo may be observed over time. Whole mount ISH to examine the effects on HOXC6 and HOXC8, as well as their potential downstream targets, may also be carried out following CRISPR-ON of lncRNA ENSGALG00000053345 to examine their interactions further. CRISPR-ON with the additional lncRNAs identified may also reveal their functional roles and allow for further characterisation.

CHAPTER 6: IMPACT AND FUTURE PROSPECTS

In synopsis, the work presented in this thesis has generated results, which have meaningfully contributed to the progression of science by beginning to further our understanding of the mechanisms governing the transition from cervical to thoracic identity. During this project, *in vitro* reporter gene analysis using a Citrine reporter with *in ovo* electroporation enabled me to show enhancer function for multiple Hox genes across the genome, primarily the HOX genes responsible for mediation of the cervical-thoracic transition, *HOXC6* and *HOXC5*.

Enhancers play a fundamental role in regulating the precise spatial and temporal gene expression patterns during embryonic development and differentiation. Many labs have focused their efforts on understanding how enhancers function, and how to identify which *cis* regulatory elements are involved in the control of a specific gene. There are many problems with attempting to identify specific enhancers, as there is no underlying sequence feature that discerns an enhancer from other non-coding regions of DNA. This project used ATAC-Sequencing data, which depicts regions of open chromatin to locate putative enhancer regions. The enhancer pipeline developed by the Münsterberg lab benefits from the utilisation of HINT-ATAC to find differential TF footprints in regions of open chromatin, allowing further insight into regulatory mechanisms.

However, more work is still needed to streamline this pipeline. The ATAC-sequencing data set, whilst useful in depicting regions of open chromatin that potentially have functional significance, cannot alone determine the presence of an enhancer. Using more of the data available from the ENCODE data base, such as methylation and histone acetylation marks, may help to identify putative enhancer regions. Marks such as H3K27Ac are present on active enhancers, and while there are no available genome wide datasets available for Histone marks in the chick, they are readily available in the human. This coupled with searching for conserved regions should enable more accurate selection of candidate enhancer peaks.

By utilising the ability to carry out heterotopic grafts in the chick embryo, as described in Chapter 2 and 4, future studies could be carried out where this experiment is repeated, and the grafted tissue somites are then isolated and re-sequenced using the same genomics approaches as carried out in this thesis. By sequencing the grafted tissue, it would be possible to identify additional ‘fixed’ genes similar to *HOXC6* that are determined prior to somite development and critical for thoracic identification.

The chick model was ideal for this study, as the Easy Culture method of ex-ovo manipulation is relatively cheap and easy. Although the chicken embryo is useful for studying early embryonic development, embryos rarely survive past HH12 when using easy culture methods (Chapman *et al.*, 2001). In order to identify putative enhancer elements that are activated in later stages of development, such as the Hox genes within this study, the *in ovo* electroporation technique (Scaal *et al.*, 2004) had to be implemented. With this method the age of the surviving embryos increases, along with average survival rates. A disadvantage is that there are no established protocols to electroporate gastrula stage embryos *in ovo*. Instead, somite-stage embryos from HH10 onwards had to be injected and targeted, for example in the neural tube or epithelial somites, which limits the area of investigation within the embryo. For further studies, it may be possible to use lipofection to deliver the enhancer-reporter construct into embryonic tissues *in-ovo*.

To our trouble, the functional analysis was interrupted due to time constraints. Therefore, the results could not provide full functional evidence for regulation of the *HOXC6* and *HOXC8* enhancers by the differential footprints that were revealed with HINT-ATAC. However, following targeted mutation of the majority of differential FP binding sites within the enhancers, and observation of a decrease in enhancer activity observed after electroporation, I believe the transcription factors responsible for *HOXC6* and *HOXC8* enhancer activity have been identified. Thus, I believe this study has provided key insights into the highly conserved region of sequence local to *HOXC6* and *HOXC8* identified as putative *cis* regulatory elements/enhancers and regulatory pathways pertinent to the function of conserved sequences in the HOXC cluster.

The novel qualities of next generation sequencing allow the transcriptome to be interrogated at great depth. In this project, new insight has also been given into signalling cascades following activation of *HOXC6*, with potential downstream targets identified through the combination of RNA-Seq and ATAC-seq analysis. However, further functional analysis is required to confirm this. Perhaps the use of CRISPR-ON techniques to overexpress *HOXC6* followed by RT-qPCR analysis would identify any changes in expression of the proposed downstream targets of *HOXC6* and thus verify their interactions. It would also be interesting to use classic gain-of-function and loss-of-function techniques to investigate the roles of these hypothesised *HOXC6* targets within cervical and thoracic development specifically.

Alongside the further development of next generation sequencing platforms, more sophisticated techniques for manipulation of RNA for sequencing are also emerging, most

notably with the advent of single-cell RNA-Seq (scRNA-Seq). Through the development of techniques for the isolation of single cells, and amplification of the subsequent extracted RNA, it has been made possible to perform sequencing on the RNA of a single cell (Hebenstreit, 2012). This provides multiple avenues for potential analyses, particularly in a developmental setting, in which there are cases where only a small number of cells convey a distinct role (Ohnishi *et al.*, 2014). Additionally, utilising scRNA-Seq would allow for purer samples for analysis. Future studies using the *HOXC6* enhancer discovered in this project as a fluorescent marker to separate out the single cells containing *HOXC6* activity for analysis would likely uncover many more regulatory genes or genetic markers, which could be utilized to fully understand the transition to specific cell types.

Future investigations could also use single cell RNA-Seq on the progenitor cells of the somites investigated in this thesis to uncover the full complexity and molecular profiles of progenitor cells that are present. This could also be linked with effects of positional identity on somitic cell fate.

Successful analysis of the expression patterns (via whole mount *in situ* hybridisation) of novel lncRNAs identified from the RNA-Seq dataset was carried out, with one lncRNA sharing an expression pattern with *HOXC6*. The use of a new CRISPR-ON technique to increase the expression of endogenous *HOXC6* across the embryo revealed a possible interaction between that novel lncRNA and *HOXC6*. This evidence that I have provided in this study lends some early support to the possibility of a more critical role for lncRNAs in governing the anterior-posterior patterning in vertebrates.

It would be useful to further understand the exact functions of all the differential lncRNAs identified in the RNA-Seq dataset. Find out the proteins, DNA or/and RNA that the lncRNAs interact with would be beneficial here. Current methods for this have been largely designed for mammalian cell culture studies and require a very large number of cells to carry this out. Designing a methodology that would allow us to investigate these interactions in the chick would be ideal. The ATAC-Seq dataset allows for identification of a putative promoter element within each of the lncRNAs and thus functional experimentation using the CRISPR-ON technique could be carried out in the future.

In conclusion, this study revealed new insights into the molecular mechanisms behind multiple Hox genes across the genome, primarily the HOX genes responsible for mediation of the cervical-thoracic transition, *HOXC6* and *HOXC5*. Novel avenues for further research have also been revealed with the discovery of novel lncRNAs that indicate a much more critical role for lncRNAs

in anterior-posterior patterning during the early stages of development. Further studies will aim to characterise more *cis*-regulatory elements, and further our understanding of their involvement in gene regulatory networks involved with somite development throughout embryogenesis. It is critical that we understand how these non-coding regions of DNA work in biological systems, to understand diseases associated with them and to potentially develop therapies that target these regions.

REFERENCES

- Altman, N. and Krzywinski, M. (2017) 'Clustering', *Nature Methods*, (14), pp. 545–546.
- Alvares, L.E., Schubert, F.R., Thorpe, C., Mootoosamy, R.C., Cheng, L., Parkyn, G., Lumsden, A. and Dietrich, S. (2003) 'Intrinsic, Hox-dependent cues determine the fate of skeletal muscle precursors', *Developmental Cell*, 5(3), pp. 379-90.
- Amano, T., Sagai, T., Tanabe, H., Mizushima, Y., Nakazawa, H. and Shiroishi, T. (2009) 'Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription', *Developmental Cell*, 16(1), pp. 47-57.
- Anders, S., Pyl, P.T. and Huber, W. (2015) 'HTSeq - a Python framework to work with high-throughput sequencing data', *Bioinformatics*, 32(2), pp. 166-169.
- Andrews, S. (2010) 'FastQC: a quality control tool for high throughput sequence data'. Available at: <http://www.bioinformatics.babraham.ac.uk/project/fastqc>.
- Aulehla, A. and Pourquié, O. (2010) 'Signaling gradients during paraxial mesoderm development', *Cold Spring Harb Perspect Biology*, 2(2).
- Baker, R.E., Schnell, S and Maini, P.K. (2006) 'A clock and wavefront mechanism for somite formation', *Developmental Biology*, 293(1), pp. 116-26.
- Baker, K.E. and Parker, R. (2004) 'Nonsense-mediated mRNA decay: terminating erroneous gene expression', *Current Opinion in Cell Biology*, 16, pp. 293-299.
- Banerji, J., Rusconi, S. and Schaffner, W. (1981) 'Expression of a β -globin gene is enhanced by remote SV40 DNA sequences', *Cell*, 27(2), pp. 299-308.
- Barolo, S. (2011) 'Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy', *BioEssays*, 34(2), pp. 135-141.

Baumgart, M., Groth, M., Priebe, S., Savino, A., Testa, G., Dix, A., Ripa, R., Spallotta, F., Gaetano, C., Ori, M., Tozzini, E.T., Guthke, R., Platzer, M. and Cellerino, A. (2014) 'RNA-Seq of the aging brain in the short-lived fish *N. furzeri*-conserved pathways and novel genes associated with neurogenesis', *Aging Cell*, 13(6), pp. 965-974.

Baykal, B. and Korkusuz, P. (2016) *Development of the Musculoskeletal System. in Musculoskeletal Research and Basic Science*. Springer International Publishing.

Bénazéraf, B., Beaupeux, M., Tchernookov, M., Wallingford, A., Salisbury, T., Shirtz, A., Shirtz, A., Huss, D., Pourquié, O., François, P. and Lansford, R. (2017) 'Multi-scale quantification of tissue behaviour during amniote embryo axis elongation', *Development*, 144(23), pp. 4462-4472.

Benjamin, A.M., Nichols, M., Burke, T.W., Ginburg, G.S. and Lucas, J.E. (2014) 'Comparing reference-based RNA-Seq mapping methods for nonhuman primate data', *BMC Genomics*, (15), pp. 570.

Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., Olenyi, T., Qiu, J., Schütze, K., Yachdav, G., Ashkenazy, H., Ben-Tal, N., Bromberg, Y., Goldberg, T., Kajan, L., O'Donoghue, S., Sander, C., Schafferhans, A., Schlessinger, A., Vriend, G., Mirdita, M., Gawron, P., Gu, W., Jarosz, Y., Trefois, C., Steinegger, M., Schneider, R. and Rost, B. (2021) 'PredictProtein – Predicting Protein Structure and Function for 29 Years', *Nucleic Acids Research*, 49(1), pp. 535-540.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. and Thurman, R.E. (2007) 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature*, 447, pp. 799-816.

Borycki, A.G., Mendham, L. and Emerson, C.P. (1998) 'Control of somite patterning by Sonic hedgehog and its downstream signal response genes', *Development*, 125, pp. 77-790.

Bond, A.M., Vangompel, M.J., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F. and Kohtz, J.D. (2009) 'Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry', *Nature Neuroscience*, 12, pp. 1020-1027.

Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H.F., Avner, P. and Ballabio, A. (1991) 'Characterization of a murine gene expressed from the inactive X chromosome', *Nature*, 351, pp. 325-329.

Borycki A., Brown A.M. and Emerson C.P. (2000) 'Shh and Wnt signaling pathways converge to control Gli gene activation in avian somites', *Development*, 127(10), pp. 2075-87.

Brent, A.E., Schweitzer, R., Fau - Tabin, C.J. and Tabin, C.J. (2003) 'A somitic compartment of tendon progenitors', *Cell*, 113, pp. 235-48.

Bridoux, L., Zarrineh, P., Mallen, J., Phuycharoen, M., Latorre, V., Ladam, F., Losa, M., Baker, S.M., Sagerstrom, C., Mace, K.A., Rattray, M. and Bobola, N. (2020) 'HOX paralogs selectively convert binding of ubiquitous transcription factors into tissue-specific patterns of enhancer activation', *PLoS Genetics*, 16(12).

Brill, G., Kahane, N., Carmeli, C., von Schack, D., Barde, Y.A. and Kalcheim, C. (1995) 'Epithelial-mesenchymal conversion of dermatome progenitors requires neural tube-derived signals: characterization of the role of Neurotrophin-3', *Development*, 121, pp. 2583-2594.

Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., McCabe, V.M., Norris, D.P., Penny, G.D., Patel, D. and Rastan, S. (1991) 'Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome', *Nature*, 351, pp. 329-331.

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S. and Rastan, S. (1992) 'The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus', *Cell*, 71, pp. 515-526.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and Willard, H.F. (1991) 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Nature*, 349, pp. 38-44.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J. and Willard, H.F. (1992) 'The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus', *Cell*, 71, pp. 527-542.

Brunner, A.L., Beck, A.H., Edris, B., Sweeney, R.T., Zhu, S.X., Li, R., Montgomery, K., Varma, S., Gilks, T., Guo, X., Foley, J.W., Witten, D.M., Giacomini, C.P., Flynn, R.A., Pollack, J.R., Tibshirani, R., Chang, H.Y., van de Rijn, M. and West, R.B. (2012) 'Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers', *Genome Biology*, 13, pp. 75.

Bulajić, M., Srivastava, D., Dasen, J.S., Wichterle, H., Mahony, S. and Mazzoni, E.O. (2020) 'Differential abilities to engage inaccessible chromatin diversify vertebrate Hox binding patterns', *Development*, 147(22), pp. 194.

Burke, A.C., Nelson, C.E., Morgan, B.A. and Tabin, C. (1995) 'Hox genes and the evolution of vertebrate axial morphology', *Development*, 121(2), pp. 333-346.

Burke, A.C. and Nowicki, J.L. (2001) 'Hox Genes and Axial Specification in Vertebrates', *American Zoologist*, 41(3), pp. 687.

Casaca, A., Novoa, A. and Mallo, M. (2016) 'Hoxb6 can interfere with somitogenesis in the posterior embryo through a mechanism independent of its rib-promoting activity', *Development*, 143(3), pp. 437-48.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) 'Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes and Development*, 25, pp. 1915-1927.

Cantile, M., Cindolo, L., Napodano, G., Altieri, V. and Cillo, C. (2003) 'Hyperexpression of locus C genes in the HOX network is strongly associated in vivo with human bladder transitional cell carcinomas', *Oncogene*, 22, pp. 6462-6468.

Carapuço, M., Nóvoa, A., Bobola, N. and Mallo, M. (2005) 'Hox genes specify vertebral types in the presomitic mesoderm', *Genes and Development*, 19(18), pp. 2116-21.

Casaca, A., Novoa, A. and Mallo, M. (2016) 'Hoxb6 can interfere with somitogenesis in the posterior embryo through a mechanism independent of its rib-promoting activity', *Development*, 143(3), pp. 437-48.

Chal, J. and Pourquie, O. (2017) 'Making muscle: skeletal myogenesis in vivo and in vitro', *Development*, 144(12), pp. 2104-2122.

Chapman, S.C., Collignon, J., Schoenwolf, G.C. and Lumsden, A. (2001) 'Improved method for chick whole-embryo culture using a filter paper carrier', *Developmental Dynamics*, 220, pp. 284–289.

Chaturvedi, A., Klionsky, N. B., Nadarajah, U. and Meyers, S. P. (2018) 'Malformed vertebrae: a clinical and imaging review', *Insights Imaging*, pp. 343-55.

Chen, J.W., Zahid S Fau - Shilts, M.H., Shilts Mh Fau - Weaver, S.J., Weaver Sj Fau - Leskowitz, R.M., Leskowitz Rm Fau - Habbsa, S., Habbsa S Fau - Aronowitz, D., Aronowitz D Fau - Rokins, K. P., Rokins Kp Fau - Chang, Y., Chang Y Fau - Pinnella, Z., Pinnella Z Fau - Holloway, L., Holloway L Fau - Mansfield, J.H. and Mansfield, J.H. (2013) 'Hoxa-5 acts in segmented somites to regulate cervical vertebral morphology', *Mechanisms of Development*, 130, pp. 226-240.

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B. and Jaenisch, R. (2013) 'Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system', *Cell Research*, 23, pp. 1163-1171.

Cho, S.G., Lee, J.W., Heo, J.S. and Kim S.Y. (2014) 'Gene expression change in human dental pulp cells exposed to a low-level toxic concentration of triethylene glycol dimethacrylate: an RNA-Seq analysis', *Basic and Clinical Pharmacology and Toxicology*, 115(3), pp. 282-90.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnar, Z. and Ponting, C.P. (2010) 'Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes', *Genome Biology*, 11, pp. 72.

Christ, B., Huang, R. and Scaal, M. (2004) 'Formation and differentiation of the avian sclerotome', *Anatomy and Embryology*, 208(5), pp. 333-350.

Christ, B., Huang, R. and Scaal, M. (2007) 'Amniote somite derivatives', *Developmental Dynamics*, 236, pp. 2382-2396.

Chu, C., Quinn, J. and Chang, H.Y. (2012) 'Chromatin isolation by RNA purification (ChIRP)', *Journal of Visualized Experiments*, 61.

Clemson, C.M., McNeil, J.A., Willard, H.F. and Lawrence, J.B. (1996) 'XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure', *Journal of Cell Biology*, 132, pp. 259-275.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solex/Illumina FASTQ variants', *Nucleic Acids Research*, 38(6), pp. 1767-1771.

Darnell, R. (2012) 'CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein', *Cold Spring Harbor Protocols*, pp. 1146-1160.

Cooke, J. and Zeeman, E.C. (1976) 'A clock and wavefront model for control of the number of repeated structures during animal morphogenesis', *Journal of Theoretical Biology*, 58(2), pp. 455-476.

Couly, G.F, Coltey, P.M. and Le Douarin, N.M. (1993) 'The triple origin of skull in higher vertebrates: a study in quail-chick chimeras', *Development*, 117(2), pp. 409–429.

Darbellay, F., Bochaton, C., Lopez-Delisle, L., Mascrez, B., Tschopp, P., Delpretti, S., Zákány, J. and Duboule, D. (2019) 'The constrained architecture of mammalian Hox gene clusters', *Proceedings of the National Academy of Sciences of the United States of America*, 116, pp. 13424 - 13433.

Dard, A., Reboulet, J., Jia, Y., Bleicher, F., Duffraisse, M., Vanaker, J.M., Forcet, C. and Merabet, S. (2018) 'Human HOX Proteins Use Diverse and Context-Dependent Motifs to Interact with TALE Class Cofactors', *Cell Reports*, 22(11), pp. 3058-3071.

Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) 'Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data', *Nature Reviews Genetics*, 14, pp. 390- 403.

Delgado, I., Giovinazzo, G. and Temiño, S. (2021) 'Control of mouse limb initiation and antero-posterior patterning by Meis transcription factors', *Nature Communications*, 12, pp. 3086.

DeRuiter, C. (2010) 'Somites: Formation and Role in Developing the Body Plan', *The Embryo Project Encyclopedia*.

Deschamps, J. and Duboule, D. (2017) 'Embryonic timing, axial stem cells, chromatin dynamics, and the Hox clock', *Genes and Development*, 31(14), pp. 1406-1416.

Dias, A.S., de Almeida, I., Belmonte, J.M., Glazier, J.A. and Stern, C.D. (2014) 'Somites without a clock', *Science*, 343(6172), pp. 791–795.

Diez del Corral, R. and Storey, K.G. (2004) 'Opposing FGF and retinoid pathways: a signalling switch that controls differentiation and patterning onset in the extending vertebrate body axis', *Bioessays*, 26(8), pp. 857-69.

Di Gesualdo, F., Capaccioli, S. and Lulli, M. (2014) 'A pathophysiological view of the long noncoding RNA world', *Oncotarget*, 5, pp. 10976-10996.

Dieci, G., Preti, M. and Montanini, B. (2009) 'Eukaryotic snoRNAs: a paradigm for gene expression flexibility', *Genomics*, 94, pp. 83-88.

Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M. and Jaffrezic, F. (2013) 'A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis', *Briefings in Bioinformatics*, 14, pp. 671–683.

Dimitrova, N., Zamudio, J.R., Jong, R.M., Soukup, D., Resnick, R., Sarma, K., Ward, A.J., Raj, A., Lee, J.T., Sharp, P.A. and Jacks, T. (2014) 'LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint', *Molecular Cell*, 54, pp. 777-790.

Dinger, M.E., Pang, K.C., Mercer, T.R. and Mattick, J.S. (2008) 'Differentiating protein-coding and noncoding RNA: challenges and ambiguities', *PLoS Computational Biology*, 4, e1000176.

Doane, A. and Elemento, O. (2017) 'Regulatory elements in molecular networks', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 9(3), p.e1374.

Duboule, D. (1998) 'Vertebrate Hox gene regulation: clustering and/or colinearity?', *Current Opinion in Genetics and Development*, 8, pp. 514-518.

Dubrulle, J. and Pourquié, O. (2004) 'Coupling segmentation to axis formation', *Development*, 131(23), pp. 5783.

Dubrulle, J. and Pourquié, O. (2004) 'fgf8 mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo', *Nature*, 427, 419–422.

Durst A. J. (2019) 'Some Questions and Answers About the Role of Hox Temporal Collinearity in Vertebrate Axial Patterning', *Frontiers in cell and developmental biology*, 7, pp. 257.

Eberwine, J., Sul, J.Y., Bartfai, T. and Kim, J. (2013) 'The promise of single-cell sequencing', *Nature Methods*, 11, pp. 25.

Engreitz, J., Lander, E.S. and Guttman, M. (2015) 'RNA antisense purification (RAP) for mapping RNA interactions with chromatin', *Methods in Molecular Biology*, 1262, pp. 183-197.

Evans, D.J.R. (2003) 'Contribution of somitic cells to the avian ribs', *Developmental Biology*, 256(1), pp. 115-127.

Falaleeva, M. and Stamm, S. (2013) 'Processing of snoRNAs as a new source of regulatory noncoding RNAs: snoRNA fragments form a new class of functional RNAs', *Bioessays*, 35, pp. 46-54.

Fatica, A. and Bozzoni, I. (2014) 'Long non-coding RNAs: new players in cell differentiation and development', *Nature Reviews Genetics*, 15, pp. 7-21.

Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P. and Kohtz, J.D. (2006) 'The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator', *Genes and Development*, 20, pp. 1470-1484.

Fonseca, N A., Rung, J., Brazma, A. and Marioni, J.C. (2012) 'Tools for mapping high-throughput sequencing data', *Bioinformatics*, 28(24), pp. 3169-3177.

Frankel, N., Davis, G., Vargas, D., Wang, S., Payre, F. and Stern, D. (2010) 'Phenotypic robustness conferred by apparently redundant transcriptional enhancers', *Nature*, 466(7305), pp. 490-493.

- Gaiti, F., Fernandez-Valverde, S.L., Nakanishi, N., Calcino, A.D., Yanai, I., Tanurdzic, M. and Degnan, B.M. (2015) 'Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity', *Molecular Biology and Evolution*, 32, pp. 2367-2382.
- Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A. and Couso, J.P. (2007) 'Peptides encoded by short ORFs control development and define a new eukaryotic gene family', *PLoS Biology*, 5, e106.
- Galis, F. (1999) 'Why do almost all mammals have seven cervical vertebrae? Developmental constraints, Hox genes, and cancer', *Journal of Experimental Zoology*, 285(1), pp. 19-26.
- Gasparini, M., Tome, J. and Shendure, J. (2020) 'Towards a comprehensive catalogue of validated and target-linked human enhancers', *Nature Reviews Genetics*, 21, pp. 292-310.
- Gaunt, S.J., Coletta, P.L., Pravtcheva, D. and Sharpe, P.T. (1990) 'Mouse Hox-3.4: homeobox sequence and embryonic expression patterns compared with other members of the Hox gene network', *Development*, 109(2), pp. 329-39.
- Gaunt, S.J. (1994) 'Conservation in the Hox code during morphological evolution', *International Journal of Developmental Biology*, 38(3), pp. 549-52.
- Gaunt, S., Sharpe, P.T. and Duboule, D. (1988) 'Spatially restricted domains of homeo-gene transcripts in mouse embryos: relation to a segmented body plan', *Development*, 104, pp. 169-179.
- Gilbert, S. F. (2000) 'Paraxial Mesoderm: The Somites and Their Derivatives'.
- Gilbert, S. F. (2010) 'Developmental Biology', 27, 9th edition. Sunderland: Sinauer Associates.
- Gilles, A.F. and Averof, M. (2014) 'Functional genetics for all: engineered nucleases, CRISPR and the gene editing revolution', *Evodevo*, 5, pp. 43.
- Giménez, C., Ielpi, M. and Mutto, A. (2016) 'CRISPR-on system for the activation of the endogenous human INS gene', *Gene Therapies*, 23, pp. 543-547.

- Gomez, C. and Pourquié, O. (2009) 'Developmental control of segment numbers in vertebrates', *Journal of Experimental Zoology, Part B, Molecular and Developmental Evolution*, 312(6), pp. 533-544.
- Gong, C. and Maquat, L.E. (2011) 'lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements', *Nature*, 470, pp. 284-288.
- Grosshans, H. and Filipowicz, W. (2008) 'Molecular biology: the expanding world of small RNAs', *Nature*, 451, pp. 414-416.
- Gouti, M., Delile, J., Stamatakis, D., Wymeersch, F.J., Huang, Y., Kleinjung, J., Wilson, V. and Briscoe, J. (2017) 'A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development', *Developmental Cell*, 41(3), pp. 243-261.
- Guibentif, C., Griffiths, J.A., Imaz-Rosshandler, I., Ghazanfar, S., Nichols, J., Wilson, V., Göttgens, B. & Marioni, J.C. (2021) 'Diverse routes towards early somites in the mouse embryo', *Developmental Cell*, 56(1), pp. 141-153
- Guillaume, R., Bressan, M. and Herzlinger, D. (2009) 'Paraxial Mesoderm Contributes Stromal Cells to the Developing Kidney', *Developmental Biology*, 329(2), pp. 169-75.
- Guillot, C., Djéffal, Y., Michaut, A., Rabe, B. and Pourquié, O. (2021) 'Dynamics of primitive streak regression controls the fate of neuromesodermal progenitors in the chicken embryo', *eLife*, 10:e64819.
- Guerreiro, I., Nunes, A., Woltering, J.M., Casaca, A., Nóvoa, A., Vinagre, T., Hunter, M.E., Duboule, D. and Mallo, M. (2013) 'Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine', *PNAS*, 110(26), 10682-6.
- Gummalla, M., Maeda, R.K., Castro Alvarez, J.J., Gyurkovics, H., Singari, S., Edwards, K.A., Karch, F. and Bender, W. (2012) 'abd-A regulation by the iab-8 noncoding RNA', *PLoS Genetics*, 8, e1002720.
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W. and Cassady, J.P. (2009) 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458, pp. 223-227.

Hainer, S.J., Pruneski, J.A., Mitchell, R.D., Monteverde, R.M. and Martens, J.A. (2011) 'Intergenic transcription causes repression by directing nucleosome assembly', *Genes and Development*, 25, pp. 29-40.

Hebenstreit, D. (2012) 'Methods, Challenges and Potentials of Single Cell RNAseq', *Biology*, 1(3), pp. 658–667.

Heo, J.B. and Sung, S. (2011) 'Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA', *Science*, 331, pp. 76-79.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P. and Ulitsky, I. (2015) 'Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species', *Cell Reports*, 11, pp. 1110-1122.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., Dodgson, J.B., Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., Fulton, R.S., Graves, T.A., Kremitzki, C., Layman, D., Magrini, V., McPherson, J.D., Miner, T.L., Minx, P., Nash, W.E., Nhan, M.N., Nelson, J.O., Oddy, L.G., Pohl, C.S., Randall-Maher, J., Smith, S.M., Wallis, J.W., Yang, S.P., Romanov, M.N., Rondelli, C.M., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H.G., Robertson, L., Masabanda, J.S., Griffin, D.K., Vignal, A., Fillon, V., Jacobsson, L., Kerje, S., Andersson, L., Crooijmans, R.P.M., Aerts, J., van der Poel, J.J., Ellegren, H., Caldwell, R.B., Hubbard, S.J., Grafham, D.V., Kierzek, A.M., McLaren, S. R., Overton, I.M., Arakawa, H., Beattie, K.J., Bezzubov, Y., Boardman, P.E., Bonfield, J.K., Croning, M.D.R., Davies, R.M., Francis, M.D., Humphray, S.J., Scott, C.E., Taylor, R.G., Tickle, C., Brown, W.R.A., Rogers, J., Buerstedde, J.M., Wilson, S.A., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M.M., Inoko, H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G. K.S., Wang, J., Liu, B., Wang, J., Yu, J., Yang, H., Nefedov, M., Koriabine, M., deJong, P.J., Goodstadt, L., Webber, C., Dickens, N.J., Letunic, I., Suyama, M., Torrents, D. and von Mering, C. (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature*, 432(7018), pp. 695-716.

Hirano, S., Hirako, R., Kajita, N. and Norita, M. (1995) 'Morphological analysis of the role of the neural tube and notochord in the development of somites', *Anatomy and Embryology*, 192(5), pp. 445-457.

Hirao, A. and Aoyama, H. (2004) 'Somite development without influence of the surface ectoderm in the chick embryo: the compartments of a somite responsible for distal rib development', *Development, Growth and Differentiation*, 46(4), pp. 351-362.

Huang, R., Zhi Q Fau - Brand-Saberi, B., Brand-Saberi B Fau - Christ, B. and Christ, B. (2000a) 'New experimental evidence for somite resegmentation', *Anatomy and Embryology*, 202, pp. 195-200.

Huang, R., Zhi, Q., and Patel, K. (2000b) 'Contribution of single somites to the skeleton and muscles of the occipital and cervical regions in avian embryos', *Anatomy and Embryology*, 202, pp. 375-383.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research*, 37(1), pp. 1-13.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T. and Rinn, J.L., (2010) 'A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response', *Cell*, 142, pp. 409-419.

Hubaud, A. and Pourquie, O. (2014) 'Signalling dynamics in vertebrate segmentation', *Nature Reviews Molecular Cell Biology*, 15(11), pp. 709-21.

Hutchinson, J.N., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B. and Chess, A. (2007) 'A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains', *BMC Genomics*, 8, pp. 39.

Imura, T., Denans, N. and Pourquie, O. (2009) 'Establishment of Hox vertebral identities in the embryonic spine precursors', *Current Topics in Developmental Biology*, 88, pp. 201-34.

Illumina Inc. (2011) 'Quality scores for next generation sequencing. Illumina Technical Note', Available at www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf. [Accessed 23 August 2022].

International Chicken Genome Sequencing Consortium (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature*, 432, pp. 695-716.

Jacob, F. and Monod, J. (1961) 'Genetic regulatory mechanisms in the synthesis of proteins', *Journal of Molecular Biology*, 3(3), pp. 318-356.

Jegalian, B.G., Miller, R.W., Wright, C.V., Blum, M. and De Robertis, E.M. (1992) 'A Hox 3.3-lacZ transgene expressed in developing limbs', *Mechanisms of Development*, 39(3), pp. 171-80.

Kadakkuzha, B.M., Liu, X.A., Swarnkar, S. and Chen, Y. (2016) 'Chapter 18 - Genomic and Proteomic Mechanisms and Models in Toxicity and Safety Evaluation of Nutraceuticals', Gupta, R. C. (ed), *Nutraceuticals*, Boston: Academic Press, pp. 227-237.

Kant, R. and Goldstein, R.S. (1999) 'Plasticity of axial identity among somites: cranial somites can generate vertebrae without expressing Hox genes appropriate to the trunk', *Developmental Biology*, 216(2), pp. 507-520.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J. and Hofacker, I.L. (2007) 'RNA maps reveal new RNA classes and a possible function for pervasive transcription Science', 316, pp. 1484-1488.

Kardong, K. (2012) 'Vertebrates: Comparative Anatomy, Function, Evolution', 6th edition.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., Regev, A., Lander, E.S. and Rinn, J.L. (2009) 'Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression', *Proceedings of the National Academy of Sciences of the United States of America*, 106, pp. 11667-11672.

Kieny, M., Mauger, A. and Sengel, P. (1972) 'Early regionalization of somitic mesoderm as studied by the development of axial skeleton of the chick embryo', *Developmental Biology*, 28(1), pp. 142-161.

Kim, W., Matsui, T., Yamao, M., Ishibashi, M., Tamada, K., Takumi, T., Kohno, K., Oba, S., Ishii, S., Sakumura, Y. and Bessho, Y. (2011) 'The period of the somite segmentation clock is sensitive to Notch activity', *Molecular Biology of the Cell*, 22(18), pp. 3541-3549.

Kim, D.H., Doyle, M.R., Sung, S. and Amasino, R.M. (2009) 'Vernalization: winter and the timing of flowering in plants', *Annual Review of Cell and Developmental Biology*, 25, pp. 277-299.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013) 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*, 14, pp. 36.

Klepikova, A.V., Logacheva, M.D., Dmitriev, S.E. and Penin, A.A. (2015) 'RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation', *BMC Genomics*, 16, pp. 466.

Kolovos, P., Knoch, T., Grosveld, F., Cook, P. and Papantonis, A. (2012) 'Enhancers and silencers: an integrated and simple model for their function', *Epigenetics and Chromatin*, 5(1), pp. 1.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P. and Pauler, F.M. (2013) 'Gene regulation by the act of long non-coding RNA transcription', *BMC Biology*, 11, pp. 59.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D.E., Zehnder, A., Qu, K., Lee, C.S., Flockhart, R.J., Groff, A.F., Chow, J., Johnston, D., Kim, G.E., Spitale, R.C., Flynn, R.A., Zheng, G.X., Aiyer, S., Raj, A., Rinn, J.L., Chang, H.Y. and Khavari, P.A. (2013) 'Control of somatic tissue differentiation by the long non-coding RNA TINCR', *Nature*, 493, pp. 231-235.

Krivega, I. and Dean, A. (2012) 'Enhancer and promoter interactions—long distance calls', *Current Opinion in Genetics and Development*, 22(2), pp. 79-85.

Krumlauf, R. (1992) 'Evolution of the Vertebrate Hox Homeobox Genes', *Bioessays*, 14, pp. 245-252.

Krumlauf R (2018) 'Hox genes, clusters and collinearity', *International Journal of Developmental Biology*, 62(11-12), pp. 659-663.

Koch, F., Scholze, M., Wittler, L., Schifferl, D., Sudheer, S., Grote, P., Timmermann, B., Macura, K. and Herrmann, B.G. (2017) 'Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors', *Developmental Cell*, 42(5), pp. 514-526.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. and Teichmann, S.A. (2015) 'The Technology and Biology of Single-Cell RNA Sequencing', *Molecular Cell*, 58(4), pp. 610-620.

- Kulesa, P.M. and Gammill, L.S. (2010) 'Neural crest migration: patterns, phases and signals', *Developmental Biology*, 344(2), pp. 566-8.
- LaRonde-LeBlanc, N.A. and Wolberger, C. (2003) 'Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior', *Genes and Development*, 17, pp. 2060-2072.
- Latos, P.A., Pauler, F.M., Koerner, M.V., Senergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E., Aumayr, K., Pasierbek, P. and Barlow, D.P. (2012) 'Airt transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing', *Science*, 338, pp. 1469-1472.
- Lee, C.S., Buttitta, L.A., May, N.R., Kripert, A. and Fan, C.M. (2000) 'SHH-N upregulates Sfrp2 to mediate its competitive interaction with WNT1 and WNT4 in the somitic mesoderm', *Development*, 127, pp. 109-118.
- Lee, C.S., Buttitta, L. and Fan, C.M. (2001) 'Evidence that the WNT-inducible growth arrest-specific gene 1 encodes an antagonist of sonic hedgehog signaling in the somite', *Proceedings of the National Academy of Sciences USA*, 98(20), pp. 11347–11352.
- Lennox, K.A. and Behlke, M.A. (2016) 'Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides', *Nucleic acids research*, 44, pp. 863-877.
- Lewis, E. (1978) 'A gene complex controlling segmentation in *Drosophila*', *Nature*, pp. 565-570.
- Li, Y., Chen, C.Y., Kaye, A.M. and Wasserman, W.W. (2015) 'The identification of cis-regulatory elements: A review from a machine learning perspective', *Biosystems*, 138, pp. 6-17.
- Li, C., Wen, A., Shen, B., Lu, J., Huang, Y. and Chang, Y. (2011) 'FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method', *BMC Biotechnology*, 12(11), pp. 92.
- Li, Z. (2019) 'Identification of transcription factor binding sites using ATAC-seq', *Genome Biology*, 20, pp. 45.

Liu, C.F. and Lefebvre, V. (2015) 'The transcription factors SOX9 and SOX5/SOX6 cooperate genome-wide through super-enhancers to drive chondrogenesis', *Nucleic Acids Research*, 43(17), pp. 8183-8203.

Louro, R., Smirnova, A.S. and Verjovski-Almeida, S. (2009) 'Long intronic noncoding RNA transcription: expression noise or expression choice?', *Genomics*, 93, pp. 291-298

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15, pp. 550.

Luo, S., Lu, J.Y., Liu, L., Yin, Y., Chen, C., Han, X., Wu, B., Xu, R., Liu, W., Yan, P., Shao, W., Lu, Z., Li, H., Na, J., Tang, F., Wang, J., Zhang, Y.E. and Shen, X. (2016) 'Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells', *Cell Stem Cell*, 18, pp. 637-652.

Ma, L., Bajic, V.B. and Zhang, Z. (2013) 'On the classification of long non-coding RNAs', *RNA Biology*, 10, pp. 925-933.

Mallo, M. (2018) 'Reassessing the Role of Hox Genes during Vertebrate Development and Evolution', *Trends in Genetics*, 34(3), pp. 209-217.

Mallo, M., Wellik, D.M. and Deschamps, J. (2010) 'Hox genes and regional patterning of the vertebrate body plan', *Developmental Biology*, 344(1), pp. 7-15.

Mallo, M. and Alonso, C.R. (2013) 'The regulation of Hox gene expression during animal development', *Development*, 140, pp. 3951-3963.

Mann, R. S. (1997) 'Why are Hox genes clustered?', *Bioessays*, 19, pp. 661-664.

Mansfield, J.H. and Abzhanov, A. (2010) 'Hox expression in the American alligator and evolution of archosaurian axial patterning', *Journal of Experimental Zoology*, 314, pp. 629-644.

Mansfield, J.H. and Abzhanov, A. (2010) 'Hox expression in the American alligator and evolution of archosaurian axial patterning', *Journal of Experimental Zoology, Biology and Molecular Developmental Evolution*, 314(8), pp. 629-644.

MacLean, J.A. (2013) 'The role of Rhox homeobox factors in tumorigenesis', *Frontiers in Bioscience*, 18(2), pp. 474–492.

Marcelle, C., Stark, M.R. and Bronner-Fraser, M. (1997) 'Coordinate actions of BMPs, Wnts, Shh and noggin mediate patterning of the dorsal somite', *Development*, 124(20), pp. 3955–3963.

Marin-Bejar, O.A.H. (2015) 'Long noncoding RNAs: from identification to functions and mechanisms', *Advances in Genomics and Genetics*, 5, pp. 257-274.

Maroto, M., Bone, R.A. and Dale, J.K. (2012) 'Somitogenesis', *Development*, 139(14), pp. 2453-2456.

Marsh, D.J., Shah, J.S. and Cole, A.J. (2014) 'Histones and their modifications in ovarian cancer - drivers of disease and therapeutic targets', *Frontiers in Oncology*, 4, pp. 144.

Marthandan, S., Priebe, S., Baumgart, M., Groth, M., Cellerino, A., Guthke, R., Hemmerich, P. and Diekmann, S. (2015) 'Similarities in Gene Expression Profiles during In Vitro Aging of Primary Human Embryonic Lung and Foreskin Fibroblasts', *Biomed Research International*, 73198.

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. and Akoulitchiev, A. (2007) 'Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript', *Nature*, 445, pp. 666-670.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet Journal*, 17, pp. 10-12.

McGinnis, W. and Krumlauf, R. (1992) 'Homeobox genes and axial patterning', *Cell*, 68(2), pp. 283-302.

McIntyre, D.C., Rakshit, S., Yallowitz, A.R., Loken, L., Jeannotte, L., Capecchi, M.R. and Wellik, D.M. (2007) 'Hox patterning of the vertebrate rib cage', *Development*, 134(16), pp. 2981.

McGrew, M.J., Sherman, A., Lillico, S.G., Ellard, F.M., Radcliffe, P.A., Gilhooly, H.J., Mitrophanous, K.A., Cambray, N., Wilson, V. and Sang, H. (2008) 'Localised axial progenitor cell populations in the avian tail bud are not committed to a posterior Hox identity', *Development*, 135(13), pp. 2289–2299.

McPherson, J.D. and Dodgson, J. (2004) 'Proposal to Sequence the Genome of The Chicken', *National Human Genome Research Institute*.

Meller, V.H., Gordadze, P.R., Park, Y., Chu, X., Stuckenholz, C., Kelley, R.L. and Kuroda, M.I. (2000) 'Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in *Drosophila*', *Current Biology*, 10, pp. 136-143.

Meller, V.H., Wu, K.H., Roman, G., Kuroda, M.I. and Davis, R.L. (1997) 'roX1 RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system', *Cell*, 88, pp. 445-457.

Mok, G.F. and Sweetman, D. (2011) 'Many routes to the same destination: lessons from skeletal muscle development', *Reproduction*, 141(3), pp. 301-12.

Mok, G.F., Folkes, L., Weldon, S.A., Maniou, E., Martinez-Heredia, V., Godden, A.M., Williams, R.M., Sauka-Spengler, T., Wheeler, G.N., Moxon, S. and Münsterberg, A.E. (2021) 'Characterising open chromatin in chick embryos identifies cis-regulatory elements important for paraxial mesoderm formation and axis extension', *Nature Communications*, 12(1), pp. 1157.

Moreau, C., Caldarelli, P., Rocancourt, D., Roussel, J., Denans, N., Pourquie, O. and Gros, J. (2019) 'Timed Collinear Activation of Hox Genes during Gastrulation Controls the Avian Forelimb Position', *Current Biology*, 29, pp. 35-50.

Moreno, E., Permanyer, J. and Martinez, P. (2011) 'The Origin of Patterning Systems in Bilateria — Insights from the Hox and ParaHox Genes in Acoelomorpha', *Genomics Proteomics Bioinformatics*, 9, pp. 65-76.

Morgan, B. A. (1996) 'Hox genes and embryonic development', *Poultry Science*, 76, pp. 96-104.

Morris, K.V. and Mattick, J.S. (2014) 'The rise of regulatory RNA', *Nature Review Genetics*, 15, pp. 423-437.

Munsterberg, A., Kitajewski, J., Bumcrot, D., McMahon, A. and Lassar, A. (1995) 'Combinatorial signaling by Sonic hedgehog and Wnt family members induces myogenic bHLH gene expression in the somite', *Genes and Development*, 9(23), pp. 2911-2922.

Nakamura, T., Largaespada, D.A., Lee, M.P., Johnson, L.A., Ohyashiki, K., Toyama, K., Chen, S.J., Willman, E.L., Chen I.M., Feinberg, A.P., Jemkins, N.A., Copeland, N.G. and Shaughnessy, J.D., Jr. (1996) 'Fusion of the nucleoporin gene NUP98 to HOXA9 by the chromosome translocation t(7;11) (P15;p15) in human myeloid leukaemia', *Nature Genetics*, 12, pp. 154-158.

NINDS (2019) 'Thoracic Outlet Syndrome Information Page | National Institute of Neurological Disorders and Stroke', Available online: <https://www.ncbi.nlm.nih.gov/pubmed/>

Nishimoto, S., Minguillon, C., Wood, S. and Logan, M.P.O. (2014) 'A Combination of Activation and Repression by a Colinear Hox Code Controls Forelimb-Restricted Expression of Tbx5 and Reveals Hox Protein Specificity', *PLOS*.

Nowicki, J.L. and Burke, A.C. (2000) 'Hox genes and morphological identity: axial versus lateral patterning in the vertebrate mesoderm', *Development*, 127(19), pp. 4265-4275.

Oberstaller, J., Joseph, S.J. and Kissinger, J.C. (2013) 'Genome-wide upstream motif analysis of *Cryptosporidium parvum* genes clustered by expression profile', *BMC Genomics*, 14, pp. 516.

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A.K., Araúzo-Bravo, M.J., Saitou, M., Hadjantonakis, A.K., Hiiragi, T. (2014) 'Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages', *Nature Cell Biology*, 16(1), pp. 27–37.

Ohno, S. (1972) 'So much "junk" DNA in our genome', *Brookhaven Symposium of Biology*, 23, pp. 366-370

Oliver, G., Wright, C.V., Hardwicke, J. and De Robertis, E.M. (1988) 'Differential antero-posterior expression of two proteins encoded by a homeobox gene in *Xenopus* and mouse embryos', *EMBO journal*, 7(10), pp. 3199-3209.

Ong, C. and Corces, V. (2011) 'Enhancer function: new insights into the regulation of tissue-specific gene expression', *Nature Reviews Genetics*, 12(4), pp. 283-293.

Ordahl, C.P. and Le Douarin, N.M. (1992) 'Two myogenic lineages within the developing somite', *Development*, 114(2), pp. 339-53.

Palmeirim, I., Henrique, D., Ish-Horowicz, D. and Pourquié, O. (1997) 'Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis', *Cell*, 91(5), pp. 639-48.

Pauli, A., Rinn, J.L. and Schier, A.F. (2011) 'Non-coding RNAs as regulators of embryogenesis', *Nature Review Genetics*, 12, pp. 136-149.

Peng, Y. and Zhang, Y. (2018) 'Enhancer and super-enhancer: Positive regulators in gene transcription', *Animal Models and Experimental Medicine*, 1(3), pp. 169-179.

Perry, M., Boettiger, A. and Levine, M. (2011) 'Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo', *Proceedings of the National Academy of Sciences*, 108(33), pp. 13570-13575.

Pettini, T. (2012) 'The role of novel long non-coding RNAs in Hox gene regulation', *FLS*. University of Manchester, Manchester.

Petruk, S., Sedkov, Y., Riley, K.M., Hodgson, J., Schweisguth, F., Hirose, S., Jaynes, J.B., Brock, H.W. and Mazo, A. (2006) 'Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference', *Cell*, 127, pp. 1209-1221.

Picelli, S., Faridani, O. and Björklund, Å. (2014) 'Full-length RNA-seq from single cells using Smart-seq2', *Nature Protocols*, 9, pp. 171–181.

Pien, S., Fleury, D., Mylne, J.S., Crevillen, P., Inze, D., Avramova, Z., Dean, C. and Grossniklaus, U. (2008) 'ARABIDOPSIS TRITHORAX1 dynamically regulates FLOWERING LOCUS C activation via histone 3 lysine 4 trimethylation', *Plant Cell*, 20, pp. 580-588.

Pineault, K.M., Song, J.Y. and Kozloff, K.M. (2019) 'Hox11 expressing regional skeletal stem cells are progenitors for osteoblasts, chondrocytes and adipocytes throughout life', *Nature Communications*, 10, pp. 3168.

Pourquie, O., Coltey, M., Breant, C. and Le Douarin, N. (1995) 'Control of somite patterning by signals from the lateral plate', *Proceedings of the National Academy of Sciences*, 92(8), pp. 3219-3223.

Prasanth, K.V. and Spector, D.L. (2007) 'Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum', *Genes and Development*, 21, pp. 11-42.

Quek, X.C., Thomson, D.W., Maag, J.L., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S. and Dinger, M.E. (2015) 'lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs', *Nucleic acids research*, 43, pp. 168-173.

Quinn, J.J. and Chang, H.Y. (2016) 'Unique features of long non-coding RNA biogenesis and function', *Nature Review Genetics*, 17, pp. 47-62.

Quinonez, S.C. and Innis, J.W. (2014) 'Human HOX gene disorders', *Molecular Genetics and Metabolism*, 111, pp. 4-15.

Rashid, D.J., Surya, K. and Chiappe, L.M. (2018) 'Avian tail ontogeny, pygostyle formation, and interpretation of juvenile Mesozoic specimens', *Scientific Reports*, 8, pp. 9014.

Rashid, D.J., Bradley, R. and Bailleul, A.M. (2020) 'Distal spinal nerve development and divergence of avian groups', *Scientific Reports*, 10, pp. 6303.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019) 'g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)', *Nucleic Acids Research*, doi:10.1093/nar/gkz369.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., Grimmond, S.M., Hume, D.A., Hayashizaki, Y. and Mattick, J.S. (2006) 'Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome', *Genome research*, 16, pp. 11-19.

Reece, J.B., Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V., Jackson, R. and Campbell, N.A. (2014), *Campbell Biology*, 10th edition. Boston: Pearson.

Remak, R. (1855) 'Untersuchungen über die Entwicklung der Wirbeltiere' Reimer, Berlin.

Reshef, R., Maroto, M. and Lassar, A.B. (1998) 'Regulation of dorsal somitic cell fates: BMPs and Noggin control the timing and pattern of myogenic regulator expression', *Genes and Development*, 12, pp. 290-303.

- Rickmann, M., Fawcett, J. W. and Keynes, R. J. (1985) 'The migration of neural crest cells and the growth of motor axons through the rostral half of the chick somite', *Journal of Embryology and Experimental Morphology*, 90, pp. 437-55.
- Rinn, J.L. and Chang, H.Y. (2012) 'Genome regulation by long noncoding RNAs', *Annual Review in Biochem*, 81, pp. 145-166.
- Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E. and Chang, H.Y. (2007) 'Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs', *Cell*, 129, pp. 1311-1323.
- Risso, D., Schwartz, K., Sherlock, G. and Dudoit, S. (2011) 'GC-Content Normalization for RNA-Seq Data', *BMC Bioinformatics*, 12, pp. 480.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139-140.
- Romanos, M., Allio, G., Roussigné, M., Combres, L., Escalas, N., Soula, C., Médevielle, F., Steventon, B., Trescases, A. and Bénazéraf, B. (2021) 'Cell-to-cell heterogeneity in Sox2 and Bra expression guides progenitor motility and destiny', *eLife*, 10:e66588.
- Scaal M. (2015) 'Early development of the vertebral column', *Seminars in Cell and Developmental Biology*, 49, pp. 83-91.
- Scaal, M. (2021) 'Development of the amniote ventrolateral body wall', *Developmental Dynamics*, 250, pp. 39– 59.
- Schmidt, M., Tanaka, M. and Münsterberg, A. (2000) 'Expression of (beta)-catenin in the developing chick myotome is regulated by myogenic signals', *Development*, 127(19), pp. 4105–4113.
- Schoenwolf, G.C. (1981) 'Morphogenetic processes involved in the remodeling of the tail region of the chick embryo', *Anatomy and Embryology*, 162, pp. 183–197.
- Scott, M. and Weiner, A. (1984) 'Structural relationships among genes that control development: Sequence homology between the Antennapedia, Ultrabithorax and fushi taraza loci of Drosophila', *Procedures of the National Academy of Science*, 81, pp. 415-4119.

Scott, M.P. (1993) 'A rational nomenclature for vertebrate homeobox (Hox) genes', *Nucleic Acids Research*, 21, pp. 1687-1688.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) 'Single-cell sequencing-based technologies will revolutionize whole-organism science', *Nature Reviews Genetics*, 14, pp. 618.

Sharpe, P.T., Miller, J.R., Evans, E.P., Burtenshaw, M.D. and Gaunt, S.J. (1988) 'Isolation and expression of a new mouse homeobox gene', *Development*, 102(2), pp. 397-407.

Shashikant, C.S., Bieberich, C.J., Belting, H.G., Wang, J.C., Borbély, M.A. and Ruddle, F.H. (1995) 'Regulation of Hoxc-8 during mouse embryonic development: identification and characterization of critical elements involved in early neural tube expression', *Development*, 121(12), pp. 4339-4347.

Smith, E., Mok, G.F. and Münsterberg, A. (2022) 'Investigating chromatin accessibility during development and differentiation by ATAC-sequencing to guide the identification of cis-regulatory elements', *Biochemical Society Transactions*, 50(3) pp. 1167-1177.

Solnica-Krezel, L. and Sepich, D.S. (2012) 'Gastrulation: making and shaping germ layers', *Annual Review Cell Developmental Biology*, 28, pp. 687-717.

Song, J.Y., Pineault, K.M., Dones, J.M., Raines, R.T. and Wellik, D.M. (2020) 'Hox genes maintain critical roles in the adult skeleton', *Proceedings of the National Academy of Science USA*, 117(13), pp. 7296-7304.

Soshnikova, N. and Duboule, D. (2009) 'Epigenetic temporal control of mouse Hox genes in vivo', *Science*, 324(5932), pp. 1320-1323.

St Laurent, G., Wahlestedt, C. and Kapranov, P. (2015) 'The Landscape of long noncoding RNA classification', *Trends in Genetics*, 31, pp. 239-251.

Stern, C.D. (2005) 'The chick; a great model system becomes even greater', *Developmental Cell*, 8, pp. 9-17.

Stern, C.D. and Piatkowska, A.M. (2015) 'Multiple roles of timing in somite formation', *Seminars in Cell and Developmental Biology*, 42, pp. 134-9.

Stem, A.M., Gall, I.C., Perry, B.L., Stimson, C.W., Weitkamp, L.R. and Poznanski, A.K. (1970) 'The hand-foot-uterus syndrome', *Journal of Pediatrics*, 77, pp. 109-116.

- Stjacques, B. and McMahon, A.P. (1996) 'Early mouse development: Lessons from gene targeting', *Current Opinion in Genetics and Development*, 6, pp. 439-444.
- Streit, A. and Stern, C.D. (1999) 'Mesoderm patterning and somite formation during node regression: differential effects of chordin and noggin', *Mechanics in Development*, 85(1-2), pp. 85-96.
- Struhl, K. (2007) 'Transcriptional noise and the fidelity of initiation by RNA polymerase II', *Nature Structures in Molecular Biology*, 14, pp. 103-105.
- Suemori, H. and Noguchi, S. (2000) 'HoxC cluster genes are dispensable for overall body plan of mouse embryonic development', *Developmental Biology*, 220, pp. 333-342.
- Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N.J. and Dean, C. (2013) 'R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus', *Science*, 340, pp. 619-621.
- Shao, P., Liao, J.Y., Guan, D.G., Yang, J.H., Zheng, L.L., Jing, Q., Zhou, H. and Qu, L.H. (2012) 'Drastic expression change of transposon-derived piRNA-like RNAs and microRNAs in early stages of chicken embryos implies a role in gastrulation', *RNA Biology*, 9(2), pp. 212-27.
- Swiezewski, S., Liu, F., Magusin, A. and Dean, C. (2009) 'Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target', *Nature*, 462, pp. 799-802.
- Tannahill, D., Cook, G.M. and Keynes, R.J. (1997) 'Axon guidance and somites', *Cell Tissue Research*, 290(2), pp. 275-83.
- Tenin, G., Wright, D., Ferjentsik, Z., Bone, R., McGrew, M. and Maroto, M. (2010) 'The chick somitogenesis oscillator is arrested before all paraxial mesoderm is segmented into somites', *BMC Developmental Biology*, 10(1), pp. 24.
- Thebault, P., Boutin, G., Bhat, W., Rufiange, A., Martens, J. and Nourani, A. (2011) 'Transcription regulation by the noncoding RNA SRG1 requires Spt2-dependent chromatin deposition in the wake of RNA polymerase II', *Molecular and cellular biology*, 31, pp. 1288-1300.
- Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E. and Chang, H.Y. (2010) 'Long noncoding RNA as modular scaffold of histone modification complexes', *Science*, 329, pp. 689-693.

- Ulitsky, I. and Bartel, D.P. (2013) 'lincRNAs: genomics, evolution, and mechanisms', *Cell*, 154, pp. 26-46.
- Van den Akker, E., Forlani, S., Chawengsaksophak, K., de Graaff, W., Beck, F., Meyer, B.I. and Deschamps, J. (2002) 'Cdx1 and Cdx2 have overlapping functions in anteroposterior patterning and posterior axis elongation', *Development*, 129(9), pp. 2181–2193.
- Vasiev, B., Balter, A., Chaplain, M., Glazier, J. A. and Weijer, C. J. (2010) 'Modelling gastrulation in the chick embryo: formation of the primitive streak', *PloS one*, 5(5), e10571-e10571.
- Veraksa, A., Del Campo, M. and McGinnis, W. (2000) 'Developmental patterning genes and their conserved functions: From model organisms to humans', *Molecular Genetics and Metabolism*, 69, pp. 85-100.
- Vinagre, T., Moncaut, N., Carapuço, M., Nóvoa, A., Bom, J. and Mallo, M. (2010) 'Evidence for a myotomal Hox/Myf cascade governing nonautonomous control of rib specification within global vertebral domains', *Developmental Cell*, 18(4), pp. 655-61.
- Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics*, 10(1), pp. 57-63.
- Weldon, S.A. and Münsterberg, A.E. (2022) 'Somite development and regionalisation of the vertebral axial skeleton', *Seminars in Cell and Developmental Biology*, 127, pp. 10-16.
- Wellik, D.M. (2007) 'Hox patterning of the vertebrate axial skeleton', *Developmental Dynamics: An Official Publication of The American Association of Anatomists*, 236(9), pp. 2454-2463.
- Wellik, D.M. and Capecchi, M.R. (2003) 'Hox10 and Hox11 genes are required to globally pattern the mammalian skeleton', *Science*, 301(5631), pp. 363-367.
- Wellik, D.M., Hawkes, P.J. and Capecchi, M.R. (2002) 'Hox11 paralogous genes are essential for metanephric kidney induction', *Genes and Development*, 16, pp. 1423- 1432.
- Williams, R.M., Candido-Ferreira, I. and Repapi, E. (2019) 'Reconstruction of the global neural crest gene regulatory network in vivo', *Developmental Cell*, 51, pp. 255–276.
- Wittkopp, P. and Kalay, G. (2011) 'Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence', *Nature Reviews Genetics*, 13(1), pp. 59-69.

Wolpert, L., Tickle, C. and Arias, A.M. (2015) 'Principles of Development', 5th edition. New York, NY, United States of America: OUP Oxford.

Woltering, J.M., Vonk, F.J., Muller, H., Bardine, N., Tuduce, I.L., de Bakker, M.A., Knochel, W., Sirbu, I.O., Durston, A.J. and Richardson, M.K. (2009) 'Axial patterning in snakes and caecilians: evidence for an alternative interpretation of the Hox code', *Developmental Biology*, 332(1), pp. 82-89.

Wu, T.D. and Watanabe, C.K. (2005) 'GMAP: a genomic mapping alignment program for mRNA and EST sequences', *Bioinformatics*, 21(5), pp. 1859-1875.

Wymeersch, F.J., Skylaki, S., Huang, Y., Watson, J.A., Economou, C., Marek-Johnston, C., Tomlinson, S.R. and Wilson, V. (2019) 'Transcriptionally dynamic progenitor populations organised around a stable niche drive axial patterning', *Development*, 146(1), dev168161.

Wymeersch, F.J., Wilson, V. and Tsakiridis, A. (2021) 'Understanding axial progenitor biology in vivo and in vitro', *Development*, 148(4), dev180612.

Xiong, F., Ma, W., Bénazéraf, B., Mahadevan, L. and Pourquié, O. (2020) 'Mechanical coupling coordinates the co-elongation of axial and paraxial tissues in avian embryos', *Development Cell*, 55, pp. 354-366.

Yang, Y., Wen, L. and Zhu, H. (2015) 'Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein', *Cell Bioscience*, 5, pp. 59.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J.C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C.G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J.G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D.N., Parker, A., Parton, A., Patricio, M., Sakthivel, M.P., Salam, A.I.A., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormass, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvilli, M., Flint, B., Frankish, A., Hunt, S.E., Ilesley, E., Kostadima, M., Langridge, N., Loveland, J.E., Marton, F.J., Morales, J., Mudge, J.M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S.J., Cunningham, F., Howe, K.L., Zerbino, D.R. and Flicek, P. (2020) 'Ensembl 2020, Nucleic Acids Research', 48(1), pp. 682–D688.

Young, J.J., Grayson, P. and Tabin, C.J. (2019) 'Developmental Biology: Hox Timing Determines Limb Placement', *Current Biology*, 29(2), pp. 52-54.

Yusuf, F. and Brand-Saberi, B. (2006) 'The eventful somite: patterning, fate determination and cell division in the somite', *Brain Structure and Function*, 211(1), pp. 21-30.

Zakany, J. and Duboule, D. (1999) 'Hox genes in digit development and evolution', *Cell and Tissue Research*, 296, pp. 19-25.

Zhao, Y. and Potter, S. (2002) 'Functional comparison of the Hoxa4, Hoxa10 and Hoxa11 Homeoboxes', *Developmental Biology*, 244, pp. 21-36.

Zimmerman, L.B., De Jesús-Escobar, J.M. and Harland, R.M. (1996) 'The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4', *Cell*, 86(4), pp. 599-606.

Zorn, A.M. and Wells, J.M. (2009) 'Vertebrate Endoderm Development and Organ Formation', *Annual Review of Cell and Developmental Biology*, 25(1), pp. 221-251.

APPENDIX 1: PRIMERS AND GUIDES

Table 1: All primer combinations used to clone cDNA fragments into pGEM-T-Easy for probe generation.

Name	Forward Primer 5'-3'	Reverse Primer 5'-3'
HOXC6	CTGGTTGCCGCTCCTATTC	TCTTGATCTGCCGTTCCGTC
HOXC8	GGAGCCATGCTCTCGTGTAT	CCCCCGGCAGCTTATCTTTG
HOXC5	CTGCCTATGCCATGCAAAGTT	AGCACAGGTTGTTGGCGAT
ENSGALG00000054525	ACTTCTTCCTCTCACATTGCC	AGAGGGAGCTTGCCTAG
ENSGALG00000050591	ACCACTCACCAGCCAAAACA	TTGTGATGTGAAGCCTCGCT
ENSGALG00000053345	CCTTTTCCTCCCCAGGGTTC	AATTTAGCCCCGGGTTGAGG
ENSGALG00000051998	ATTTCCAGAGAGCGTGGCAA	AGCTCAACCAGAGTGATGCC
ENSGALG00000047842	GCCCTTCTAATATGGCCGGT	AGGGAAGTTGCTCCCCTATTG
ENSGALG00000054479	CGTTGACTCTCAAAGCCCCT	CTCCCAGGAGTGCTCCTACT
ENSGALG00000050453	CCACTTCTGCCTGGACGATT	ACTGTTGACCGCATCCCATT
ENSGALG00000052325	ACCCAAGGCATCACTGTAGC	ACAGGATAGGACAGTCCCCC
ENSGALG00000051195	ACATTGAGTCAGGCACAGCA	CATTCCAGCCCCATGCACTA
ENSGALG00000054927	TGGTGGGTCCATTCTTCTGC	TCTCCATGTGTCACTCGCAC

Table 2: All primer combinations used to clone gDNA fragments into pTk-Citrine reporter vector for putative enhancer plasmid generation.

Name	Coordinates	Forward Primer 5'-3'	Reverse Primer 5'-3'	Signal?
HOXC8	chr33:1040749-1041457	TTTTTTCGTCTCgccag gCGCTGCGATTTCGG	TTTTTTCGTCTCcaacag TTCTCCCTCCTTCCGTT TACA	Y
HOXC6	chr33:1,053,022-1,053,708	TTTTTTCGTCTCgccag gCTCAGTGCCGAAAA TAACCAC	TTTTTTCGTCTCcaacag CCCCCTCCAAAATGC	N
HOXC6	chr33:1,054,949-1,055,690	TTTTTTCGTCTCgccag gGGCCGGGGGGG	TTTTTTCGTCTCcaacag TGCTTGGGGAAAAGG GT	N
HOXC6	chr33:1,070,802-1,071,482	TTTTTTCGTCTCgccag gAGGCGCTGCGCT	TTTTTTCGTCTCcaacag CCAATCACCACCACT CC	N
HOXC5	chr33:1,073,437-1,073,992	TTTTTTCGTCTCgccag gGAGAAAAAAAAG AGGGGATGCG	TTTTTTCGTCTCcaacag GGGGATTCTTTTTTTT TTTGCTTTTT	N
HOXC5	chr33:1,079,305-1,080,428	TTTTTTCGTCTCgccag gTATCCCCCTCTCG GCT	TTTTTTCGTCTCcaacag TGCAGCAGCCCCGG	N
HOXC5	chr33:1,089,358-1,090,015	TTTTTTCGTCTCgccag gCTCAGAGCGATCCC CG	TTTTTTCGTCTCcaacag ATATTGCCCCCTTCGCT TTC	Y
HOXC6	chr33:1053137-1053692	TTTTTTCGTCTCgccag gGGACACAAAAAGG GGACAG	TTTTTTCGTCTCcaacag GCACATTTCCCTCCGC	N
HOXC6	chr33:7,621,860-7,622,199	TTTTTTCGTCTCgccag gGGAATTGTATAAAC ATGGCGACA	TTTTTTCGTCTCcaacag CCCTCTCCTCCGAAAA AAAAAA	Y
HOXC6	chr33:1059415-1060642	TTTTTTCGTCTCgccag gTGTGTTGAGGTCAG GGGT	TTTTTTCGTCTCcaacag ACCGATAATTTTACAC AAAGTGC	N
HOXC6	chr33:1065227-1065760	TTTTTTCGTCTCgccag gACTGATGGAAAATA TCAATGGGG	TTTTTTCGTCTCcaacag GATTTTATCACTCTGTT TTTGCTCTG	N
HOXC8	chr33:7606832-7607274	TTTTTTCGTCTCgccag gAGGCACTTTCGGAG CC	TTTTTTCGTCTCcaacag GTTGTGTTGAGGGGG TGT	N
HOXC8	chr33:7617813-7618796	TTTTTTCGTCTCgccag gCTGCTACCACTATT GTTATCACCG	TTTTTTCGTCTCcaacag ATCCGGACTGTAAATG CGG	N
HOXC5	chr33:7,648,152-7,648,947	TTTTTTCGTCTCgccag gGCACGCATGCGTGC A	TTTTTTCGTCTcaacagT GATATCATCTGCGTAG GAGCC	N

HOXC8	chr33:7,615,793 -7,616,487	TTTTTTCGTCTCgccag gTTTGAACAGCTTTG ACTTTGCCT	TTTTTTCGTCTcaacagG GCTCTCTCTTTAAATA AATGCG	N
HOXC8	chr33:7,606,138 -7,606,749	TTTTTTCGTCTCgccag gCCCTCGTCCCGGT	TTTTTTCGTCTcaacagG AGGAGGTGGGCGAG	N
HOXC8	chr33:7,606,842 -7,607,412	TTTTTTCGTCTCgccag gGGAGCCGGCTCTCT G	TTTTTTCGTCTcaacagC CGGGGAAGGCACAC	N
HOXC6/8	chr33:7,617,319 -7,617,835	TTTTTTCGTCTCgccag gAAGATAATCAAAAT GCTTCGATCG	TTTTTTCGTCTcaacagA TAAGAGCAAACGCGG T	N
HOXC6/8	chr33:7,617,861 -7,618,837	TTTTTTCGTCTCgccag gATCACCCTATTTTT CGTTTTG	TTTTTTCGTCTcaacagC ATCACCCATCGTATGA CTG	N
HOXC6	chr33:7,621,723 -7,622,272	TTTTTTCGTCTCgccag gAAAAAAAAAAAAAGA AGGGGGGGGAG	TTTTTTCGTCTcaacagG CCCCGAGGAGAGAGA	N
HOXC6	chr33:7,622,469 -7,623,513	TTTTTTCGTCTCgccag gATGAGGGGCACCA GC	TTTTTTCGTCTcaacagG CATCACCGCGTCTACC	N
HOXC6	chr33:7,623,705 -7,624,229	TTTTTTCGTCTCgccag gTCCCCACCTACCTCT CCA	TTTTTTCGTCTcaacagT GAAATAATTCATTTTTA AAAGCCCAACGA	N
HOXC6	chr33:7,625,526 -7,626,039	TTTTTTCGTCTCgccag gCGCAACAGTGGGC G	TTTTTTCGTCTcaacagA TCGCCCTGACACTCG	N
HOXC6	chr33:7,628,238 -7,629,363	TTTTTTCGTCTCgccag gCAAACCTCTTCTTT CCCCAT	TTTTTTCGTCTcaacagT ATGGATACAGATGTGT GTTCTACAC	N
HOXC6	chr33:7,630,967 -7,631,526	TTTTTTCGTCTCgccag gTGCTCTGATGATGT GTGACC	TTTTTTCGTCTcaacagG CGAATTGATGAAGTGA TTTAGGA	N
HOXC6	chr33:7,632,072 -7,632,661	TTTTTTCGTCTCgccag gGAAATGTATATTAT ATGGGACGTG	TTTTTTCGTCTcaacagA GAGCGGTGCCACAC	N
HOXC6	chr33:7,635,342 -7,635,884	TTTTTTCGTCTCgccag gTGTTGTGTTTAGCTC CAGCCG	TTTTTTCGTCTcaacagT GCAGCACAGCCACAC	N
Hox c cluster 1	chr33:7,590,915 -7,592,005	TTTTTTCGTCTCgccag gCGCTGGAGGTGGA TTTTATG	TTTTTTCGTCTCcaacag CCATTGCAACGTGCTC AA	N
Hox c cluster 2	chr33:7,591,937 -7,592,618	TTTTTTCGTCTCgccag gCTTTTTTCCCCCCA AAGC	TTTTTTCGTCTCcaacag AGGGTAGGAAAAAAG CCG	N
Hox c cluster 3	chr33:7,592,778 -7,594,072	TTTTTTCGTCTCgccag gCCACCACCGATCGC T	TTTTTTCGTCTCcaacag GCAAAGGGACGCTTA AAGG	N

Hox c cluster 4	chr33:7,595,163 -7,595,776	TTTTTCGTCTCgccag gCTCTCTCCCGCTTTA TGTCG	TTTTTCGTCTCcaacag AATCCTATTGTCTGGC CG	N
Hox c cluster 5	chr33:7,599,888 -7,600,569	TTTTTCGTCTCgccag gGAAATCGCCCAACT CCATTC	TTTTTCGTCTCcaacag TCTTCCGGAAGGGGG AGG	N
Hox a cluster 1	chr2:32,648,959 -32,649,618	TTTTTCGTCTCgccag gGAAAGCCAAACCCG ACC	TTTTTCGTCTCcaacag GGCGTGGGTAGAGCT C	N
Hox a cluster 2	chr2:32,655,690 -32,656,560	TTTTTCGTCTCgccag gAGAGAGAAGGATA CGCAGATG	TTTTTCGTCTCcaacag GCAGGCTGTCTGGGC	N
Hox a cluster 3	chr2:32,657,737 -32,658,402	TTTTTCGTCTCgccag gGCGGCGTGCGT	TTTTTCGTCTCcaacag CTCAGCGTGAAAGTAT TTCTAGG	N
Hox a cluster 4	chr2:32,658,216 -32,659,086	TTTTTCGTCTCgccag gAATTGCTCCAAAGA AGG	TTTTTCGTCTCcaacag GCATACCACATCCAAA CCC	N
Hox a cluster 5	chr2:32,659,523 -32,660,306	TTTTTCGTCTCgccag gCGATAGGTTCCGAT AATCATCTGA	TTTTTCGTCTCcaacag ATGGCAACCCAGACTA CAA	N
Hox a cluster 6	chr2:32,665,925 -32,666,660	TTTTTCGTCTCgccag gATGTCATAGGGCGA GAAGAC	TTTTTCGTCTCcaacag CTAGAAGGAGTCGTTG CTTC	N
Hox a cluster 7	chr2:32,669,315 -32,670,512	TTTTTCGTCTCgccag gTCTCTGCTTTATTTG CTGTAAAGTGA	TTTTTCGTCTCcaacag TGTTGGTGACGGCGGT	N
Hox b cluster 1	chr27:6,159,930 -6,160,855	TTTTTCGTCTCgccag gAGCTGTAAGAGGT GCATCTC	TTTTTCGTCTCcaacag AGGAAGGAAGGAAAG GAAGG	N
Hox b cluster 2	chr27:6,165,427 -6,166,690	TTTTTCGTCTCgccag gCCTCAATTCCCATTC TGCAG	TTTTTCGTCTCcaacag TGTGCTAAAGATGAAT CCATGGC	N
Hox b cluster 3	chr27:6,175,336 -6,175,981	TTTTTCGTCTCgccag gATGTTTGGTTGCGT TGTTG	TTTTTCGTCTCcaacag CCATTTTTGGTATGGG GGC	N
Hox b cluster 4	chr27:6,175,974 -6,176,571	TTTTTCGTCTCgccag gAGTAAACGCGTCGG AGT	TTTTTCGTCTCcaacag GCAGAACGGACACCTC G	N
Hox b cluster 5	chr27:6,243,493 -6,244,467	TTTTTCGTCTCgccag gCGCTCTTCGCCCG	TTTTTCGTCTCcaacag ATTGGAGGAGGATGG TAACAG	N
Hox b cluster 6	chr27:6,276,163 -6,276,920	TTTTTCGTCTCgccag gGTGATTTCAGAGAT ATCCATGACC	TTTTTCGTCTCcaacag CTGCCTGGTGCTGCT	N
Hox d cluster 1	chr7:16,251,001 -16,252,025	TTTTTCGTCTCgccag gCATTGGGGTTCTT CTCCT	TTTTTCGTCTCcaacag GCCTACCGACTATTGC TCA	N

Hox d cluster 2	chr7:16,282,720-16,283,491	TTTTTTCGTCTCgccag gGGCAGAGCCCGGG	TTTTTTCGTCTCcaacag ATTGCGGCAGCCTTC	N
Hox d cluster 3	chr7:16,302,280-16,303,391	TTTTTTCGTCTCgccag gCATATTCACTTGCA GCGCTAA	TTTTTTCGTCTCcaacag CAAAAATGCGCTTTGC CTC	N
Hox d cluster 4	chr7:16,303,233-16,304,423	TTTTTTCGTCTCgccag gCCACCACCACCACC AG	TTTTTTCGTCTCcaacag AGAACCCCGAGGTTCC	N
Hox d cluster 5	chr7:16,304,344-16,305,455	TTTTTTCGTCTCgccag gAGTGGCGTAGGTCT GG	TTTTTTCGTCTCcaacag TACCAAGTCGCCCAAG AATG	N
HOXC8	chr33:7607907-7608062	TTTTTTCGTCTCgccag gTTTCCGCAGCCAAA AAAATGC	TTTTTTCGTCTCcaacag TTCAAAGAGAAGAGC ACCCACC	N
HOXC6	chr33:7621677-7622296	TTTTTTCGTCTCgccag gTGATGGTTTTGACG CGATCG	TTTTTTCGTCTCcaacag TGCCCAAAGCAACCGA GA	N
HOXC6/8	chr33:7,617,230-7,618,956	TTTTTTCGTCTCgccag gCTCCCCCACTCCCG	TTTTTTCGTCTcaacagC CCCCCACCCTTTTC	N

Table 3: All primer combinations used to clone primer overhangs for mutagenesis.

Name	Forward Primer 5'-3'	Reverse Primer 5'-3'
HOXC6 Mutation A	GAAAATTTACAGAGTCTACCGC CCCTGGGACGACTGAATCAC	CAGTCGTCCCAGGGGCGGTAG ACTCTGTAAATTTCTGC
HOXC6 Mutation B	GCTGAGCAATCCCCTAAACATC AGTCCGCACAAGTTGG	GTGATTCAGTCGTCCCAGGGG CGGTAGACTCTGTAAATTT
HOXC6 Mutation C	CATGAACGTTCCAGGGGCGGT GTTGGGCTCTTATGACTCTTT	AGTCATAAGAGCCCAACACCG CCCCCTGGAACGTTTCATGGAT
HOXC8 Mutation A	AAATGCCGCggggcaAGCTCTGT TTGTCTCTCTGC	AACAGAGCTtgccccGCGGCAT TTTTTTGGCTGC
HOXC8 Mutation B	TTTACAGCTagtgggGTCTCTCTG CTATGCGGGGC	CAGAGAGACcccactAGCTGTA AAAGCGGCATTTTTTTTGG
HOXC8 Mutation C	CGTCTGGCTggccggGCTTTATGG TTTAAATAAGGTGGGT	CCATAAAGCccggccAGCCAGA CGTCTAGCCAGATC
HOXC8 Mutation D	CTTTAATTGagggcgGGTTTAAAT AAGGTGGGTGCT	ATTTAAACCcgccctCAATTAAA GCCAGACGTCTAGCCAGAT
HOXC8 Mutation E	TGCTCTTCgagggtcAACCGGATT ATTGGAATGTTTTGTCTACA	ATCCGGTTgaccctcGAAGAGCA CCCACCTTATTTAAACC

Table 4: All guide RNAs designed for CRISPR-ON.

Name	Forward Primer 5'-3'	Reverse Primer 5'-3'
HOXC6 sg1	TTTCGTCTCcagtcGGGAGATACCA TCCTACCCCgtttaGAGACGAAA	TTTCGTCTCtaaacGGGGTAGGATG GTATCTCCCgactgGAGACGAAA
HOXC6 sg2	TTTCGTCTCcagtcGGAGTCCACAT GACTTCCTGgtttaGAGACGAAA	TTTCGTCTCtaaacCAGGAAGTCAT GTGGACTCCgactgGAGACGAAA
HOXC6 sg3	TTTCGTCTCcagtcGGCACTGCTTG GACTGTAAgtttaGAGACGAAA	TTTCGTCTCtaaacTTAACAGTCCAA GCAGTGCCgactgGAGACGAAA
HOXC6 sg4	TTTCGTCTCcagtcGACTGAATCAC AAGTTGGATgtttaGAGACGAAA	TTTCGTCTCtaaacATCCAACCTTG ATTCAGTCgactgGAGACGAAA
HOXC8 sg1	TTTCGTCTCcagtcGTACTAATTTAT GGCCCCGgtttaGAGACGAAA	TTTCGTCTCtaaacCGGGGGCCATA AATTAGTACgactgGAGACGAAA
HOXC8 sg2	TTTCGTCTCcagtcGTAATAATAAC GGCGATAATgtttaGAGACGAAA	TTTCGTCTCtaaacATTATCGCCGTT ATTATTACgactgGAGACGAAA
HOXC8 sg3	TTTCGTCTCcagtcGATAGGGGGT ACATTAATAGgtttaGAGACGAAA	TTTCGTCTCtaaacCTATTAATGTAC CCCCTATCgactgGAGACGAAA
HOXC8 sg4	TTTCGTCTCcagtcGGGCGCGCGT GTCCGCCCCAggtttaGAGACGAAA	TTTCGTCTCtaaacTGGGGCGGACA CGCGCGCCCgactgGAGACGAAA

APPENDIX 2: SEQUENCES

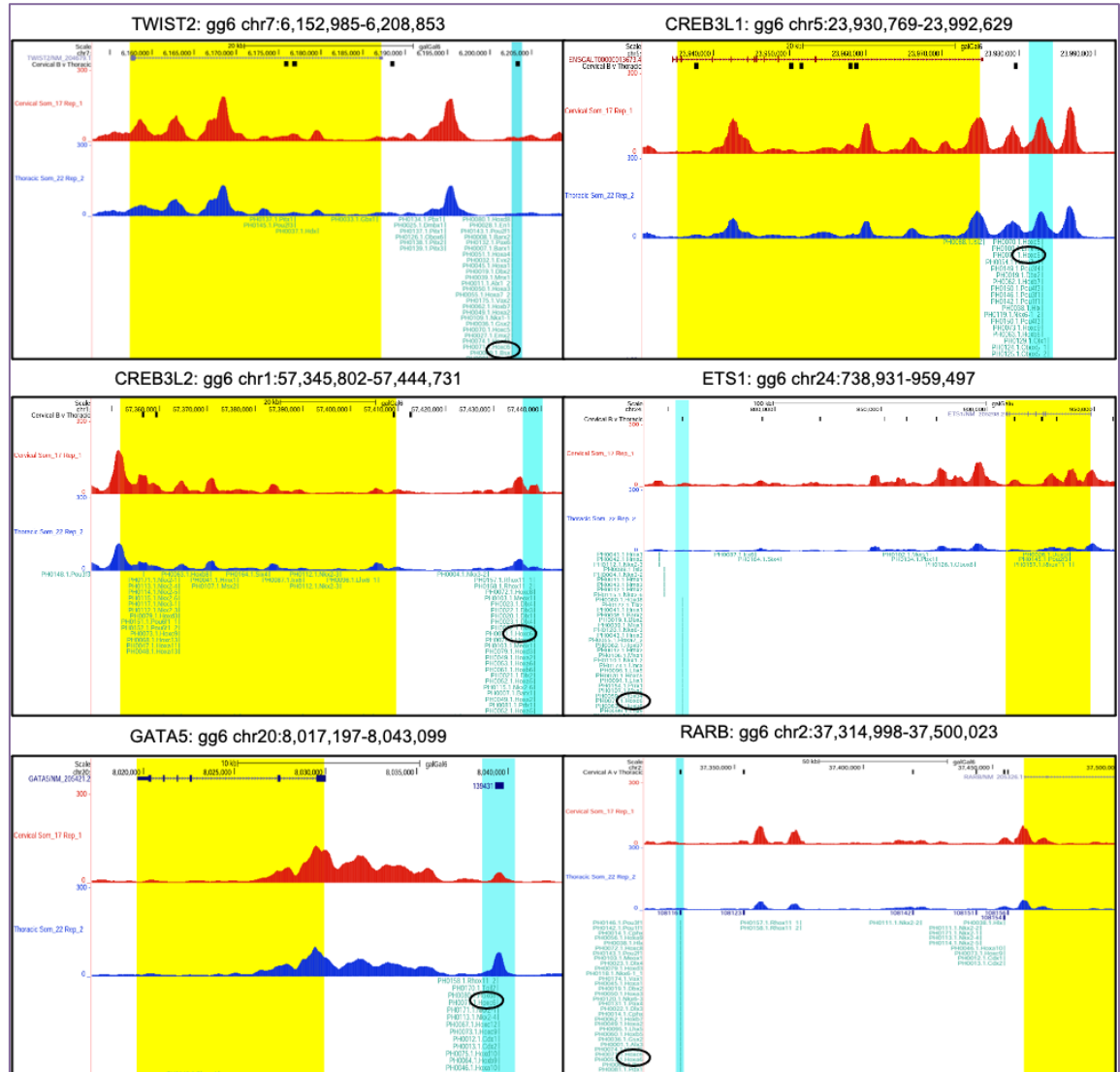
Table 5: Enhancer sequence for HOXC8. Enhancer sequence is highlighted in green. Elements selected for mutagenesis are highlighted in grey.



Table 6: Enhancer sequence for HOXC6. Enhancer sequence is highlighted in green. Elements selected for mutagenesis are highlighted in grey.

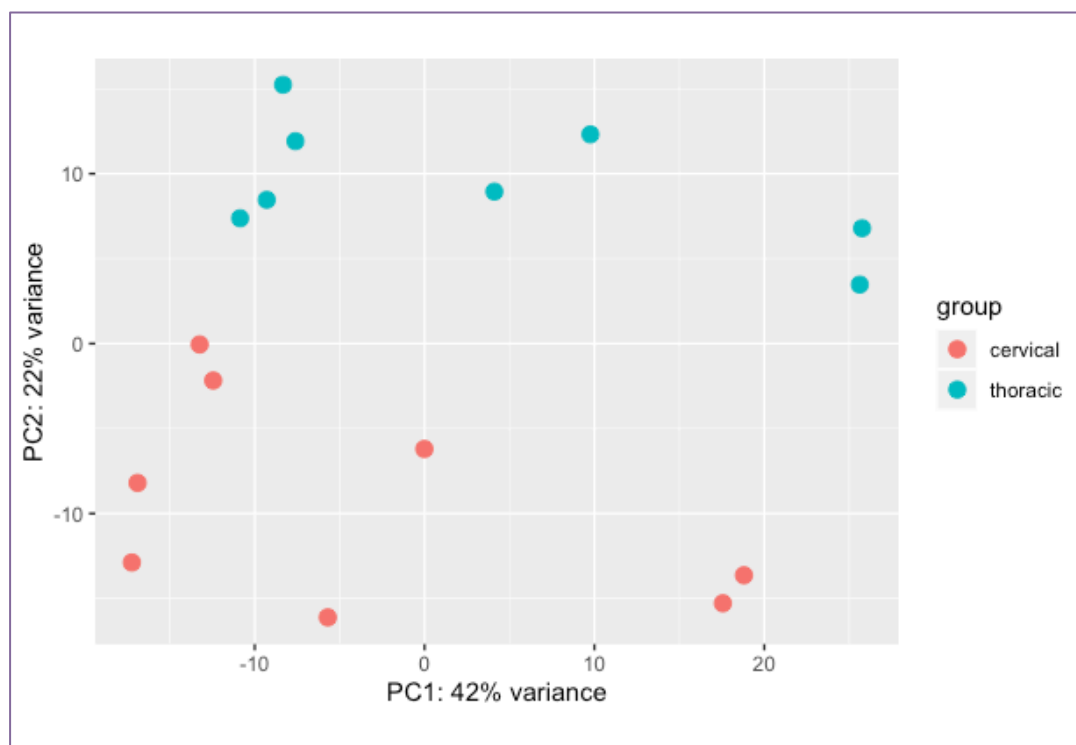


Table 7: Snapshots of the UCSC genome browser view of the 6 additional TFs that are predicted to be downstream targets of HOXC6. The gene is highlighted in yellow and the predicted CRE is highlighted in blue. HOXC6 differential footprints are circled in black within the predicted CRE.

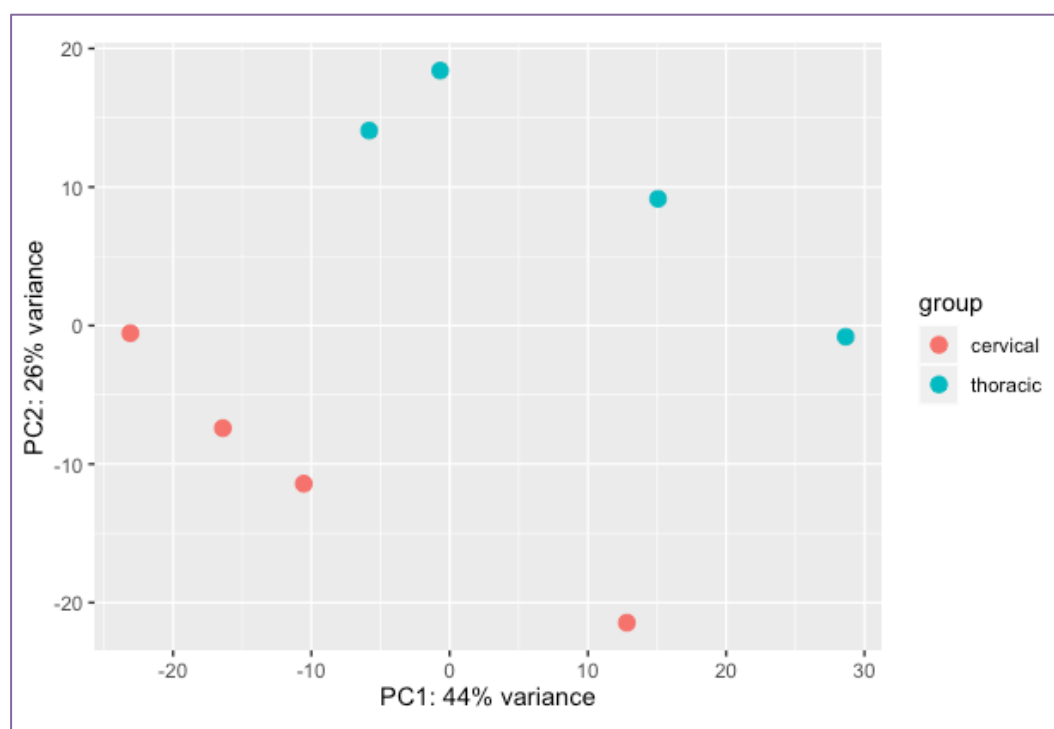


APPENDIX 3: RNA-SEQ

3.1. PCA plots



Appendix Figure 1: PCA plot of somite samples 17 and 18 (cervical – red) vs. somite samples 21 and 22 (thoracic – blue) from Differential 1 of RNA-Seq dataset.



Appendix Figure 2: PCA plot of somite samples 17 (cervical – red) vs. somite samples 22 (thoracic – blue) from Differential 1 of RNA-Seq dataset.

3.2. List of differentially expressed genes

Table 3.2.8: List of 153 upregulated genes in thoracic somites vs. cervical somites from the overlap of D1 and D2, excluding D3, from the RNA-Seq dataset

Gene ID	Gene Name	Base Mean	Log2 Fold Change	lfcSE	stat	pvalue	padj
ENSGALG0000037262	HOXC8	23.5700724	5.41126962	1.05822892	5.11351516	3.16E-07	5.38E-05
ENSGALG0000013372	IL7R	3.39000491	5.30538979	1.66766315	3.18133179	0.001466	0.03541858
ENSGALG0000039051	HOXC6	3.67630679	2.36934249	1.04314744	2.27134	0.0231264	0.0647718
ENSGALG0000054994	-	2.84823771	5.04676386	1.55845549	3.23831121	0.0012024	0.03123058
ENSGALG0000054525	-	5.69292064	4.56929844	1.11239164	4.10763464	4.00E-05	0.00284617
ENSGALG0000051998	-	2.07852157	4.33347914	1.56795082	2.76378512	0.00571352	0.08409879
ENSGALG0000047842	-	2.79263897	3.90711002	1.31454786	2.97220826	0.00295666	0.05563497
ENSGALG0000001252	CREB3L3	3.65995759	3.81628746	1.1237901	3.39590771	0.00068401	0.02139987
ENSGALG0000003424	LMX1A	6.14168907	3.79446437	0.98422229	3.85529205	0.00011559	0.0061425
ENSGALG0000054479	-	2.47261494	3.25147236	1.1521606	2.82206523	0.00477155	0.0747215
ENSGALG0000005790	DEPTOR	4.72179832	3.12567333	1.07100297	2.91845442	0.00351771	0.06292701
ENSGALG0000050453	-	4.53280993	3.07060846	0.99986845	3.07101244	0.00213334	0.04468889
ENSGALG0000042705	KCNK3	8.18741614	2.8219823	0.91383264	3.08807343	0.00201459	0.04313981
ENSGALG0000052325	-	7.89506036	2.75447265	0.95602167	2.88118224	0.00396187	0.06752924
ENSGALG0000008947	PAK5	20.4742209	2.65425561	0.77129093	3.44131572	0.00057889	0.01919109
ENSGALG0000006323	LECT2	2.87069072	2.64291088	0.9212175	2.86893255	0.0041186	0.06910479

ENSGALG00 000051195	-	7.6317 3914	2.5884 6848				0.0814 13
ENSGALG00 000009509	CALCR	10.798 5976	2.4424 2332				0.0917 2091
ENSGALG00 000053595	CXXC4	12.988 4117	2.3470 0143				0.0008 8697
ENSGALG00 000002203	FGF18	48.089 8762	2.1920 705				0.0023 7224
ENSGALG00 000043764	KCNS2	11.012 4828	2.1829 5503				0.0977 4904
ENSGALG00 000054927	-	9.3871 8406	2.1047 8301				0.0964 3114
ENSGALG00 000054348	-	7.6082 2727	2.0344 6224				0.0840 9879
ENSGALG00 000016384	MYT1L	20.486 4834	1.9502 2425				0.0565 0674
ENSGALG00 000035943	GSX2	18.811 2082	1.9184 7224				0.0390 289
ENSGALG00 000048420	-	17.548 387	1.6773 9173				0.0913 8964
ENSGALG00 000015846	SNAP91	45.379 3923	1.6318 6143				0.0037 6347
ENSGALG00 000051874	-	15.335 9726	1.6044 1138				0.0312 3058
ENSGALG00 000033365	ALDH1A3	284.21 2621	1.5567 8717				1.78E- 05
ENSGALG00 000049111	-	14.428 8605	1.5372 4653				0.0523 212
ENSGALG00 000009006	STEAP1	37.724 0895	1.4709 1195				0.0120 3401
ENSGALG00 000016281	DMD	1611.3 2692	1.4679 3378				3.75E- 15
ENSGALG00 000007668	BMP7	42.737 626	1.4607 1421				0.0102 4533
ENSGALG00 000015799	ELAVL2	85.052 0638	1.4556 8711				0.0001 2366
ENSGALG00 000017155	C11orf87	23.795 2973	1.4285 8308				0.0332 5536

ENSGALG00 000032410	CBFA2T3	19.106 4694	1.3459 2486				0.0330 4708
ENSGALG00 000010160	DMRT1	36.183 5104	1.3091 1043	0.41932026 0.4137664	3.20977776 3.16388774	0.00132838 0.00155677	0.0369 4892
ENSGALG00 000048001	-	48.932 6658	1.2978 4788	0.38712127	3.35256158	0.00080067	0.0236 1826
ENSGALG00 000012023	DACT1	76.267 4586	1.2631 9	0.39684926	3.18304739	0.00145734	0.0353 5062
ENSGALG00 000010738	TC2N	63.399 0346	1.2599 3958	0.39384729	3.19905608	0.00137878	0.0337 207
ENSGALG00 000004853	LRP3	83.905 4693	1.1690 399	0.29316572	3.98764189	6.67E-05	0.0040 3583
ENSGALG00 000002993	CA10	76.359 9191	1.1583 5168	0.31713255	3.65257899	0.00025962	0.0110 3689
ENSGALG00 000027533	RNF24	63.396 4315	1.1549 1389	0.29297048	3.94208284	8.08E-05	0.0046 8896
ENSGALG00 000007945	CRYAB	61.033 8019	1.1241 8073	0.34895758	3.22153983	0.00127504	0.0325 2918
ENSGALG00 000038923	-	43.883 4246	1.1129 3689	0.28665785	3.88245738	0.00010341	0.0056 0663
ENSGALG00 000036191	-	65.009 8267	1.1047 5722	0.36510284	3.02587959	0.00247911	0.0493 5683
ENSGALG00 000014743	RNF180	89.781 398	1.1018 1774	0.24805921	4.44175293	8.92E-06	0.0008 4295
ENSGALG00 000009885	HELB	49.448 6803	1.0717 9318	0.23825564	4.4985007	6.84E-06	0.0006 8691
ENSGALG00 000026055	PALM	42.382 2501	1.0188 6095	0.36451579	2.79510788	0.00518824	0.0788 547
ENSGALG00 000037129	FAM241A	29.947 2063	1.0053 6065	0.32153434	3.12675983	0.00176744	0.0394 8472
ENSGALG00 000007706	FGF8	157.82 5612	0.9947 7819	0.17034778	5.83968981	5.23E-09	1.80E- 06
ENSGALG00 000050863	-	35.947 6005	0.9781 7436	0.33976937	2.87893625	0.00399019	0.0677 7227
ENSGALG00 000005078	NUP210	100.80 696	0.9759 5957	0.23286107	4.19116669	2.78E-05	0.0021 5663
ENSGALG00 000053846	-	27.182 4068	0.9729 9572	0.33462813	2.90769255	0.00364106	0.0644 9497

ENSGALG00 000006445	ARNT2	97.183 2246	0.9679 1879	0.22523991	4.29727922	1.73E-05	0.0014 4463
ENSGALG00 000036775	SETBP1	854.81 3677	0.9654 1377	0.18959809	5.09189607	3.55E-07	5.82E- 05
ENSGALG00 000048409	TXNIP	48.522 495	0.9626 6689	0.3388728	2.84079124	0.00450018	0.0724 1396
ENSGALG00 000011623	ADAMTS3	439.97 751	0.9134 1054	0.20185885	4.52499635	6.04E-06	0.0006 3259
ENSGALG00 000005499	IL17RD	242.75 5538	0.9101 8103	0.19821932	4.59178763	4.39E-06	0.0004 776
ENSGALG00 000017031	-	66.906 5293	0.9033 9823	0.28121708	3.21245862	0.00131604	0.0329 529
ENSGALG00 000006638	GRIK4	118.06 812	0.8863 1126	0.23990417	3.69443878	0.00022037	0.0098 9215
ENSGALG00 000033997	ZNF800	106.82 8731	0.8764 9287	0.21129565	4.14818231	3.35E-05	0.0024 9686
ENSGALG00 000041263	DCLK1	38.182 0734	0.8734 6258	0.31221914	2.79759461	0.00514847	0.0785 8075
ENSGALG00 000035138	UBL5	559.59 4065	0.8714 5325	0.20240292	4.30553687	1.67E-05	0.0014 0805
ENSGALG00 000004276	CHST12	133.11 2029	0.8477 4192	0.2476991	3.42246672	0.00062056	0.0200 6606
ENSGALG00 000008951	MIGA1	36.079 0982	0.8179 3132	0.29779709	2.74660613	0.00602154	0.0863 9622
ENSGALG00 000016906	SPRY2	158.70 2622	0.8057 0653	0.19017992	4.23654886	2.27E-05	0.0017 9606
ENSGALG00 000044299	KIAA1257	54.194 2472	0.7884 572	0.27974088	2.81852691	0.00482446	0.0752 1896
ENSGALG00 000013135	GALNT1	2034.7 6867	0.7834 385	0.18786795	4.17015516	3.04E-05	0.0023 1563
ENSGALG00 000003932	SDC4	526.73 711	0.7644 3556	0.14552527	5.25294033	1.50E-07	3.01E- 05
ENSGALG00 000009530	PDZRN4	168.35 9426	0.7560 433	0.23365005	3.23579337	0.00121305	0.0313 6574
ENSGALG00 000021611	-	648.65 6872	0.7515 6604	0.1267392	5.9300205	3.03E-09	1.16E- 06
ENSGALG00 000011241	RFTN1	208.46 3358	0.7280 6256	0.19114355	3.80898317	0.00013954	0.0069 789

ENSGALG00 000002517	TMEM35B	294.93 9355	0.7162 9087				0.0027 4428
ENSGALG00 000007809	TSHZ2	514.89 4271	0.7154 9998	0.1739274	4.11833243	3.82E-05	0.0584 6449
ENSGALG00 000035031	HEY1	91.907 9254	0.6947 8582	0.20716409	3.35379464	0.00079712	0.0236 1826
ENSGALG00 000048930	CDC37	56.052 5188	0.6814 6379	0.2162868	3.15074147	0.00162857	0.0375 8686
ENSGALG00 000015842	TIAM1	186.27 6517	0.6766 4727	0.24167498	2.79982341	0.00511306	0.0783 7129
ENSGALG00 000003575	-	234.76 8234	0.6753 5448	0.1719376	3.92790454	8.57E-05	0.0049 1489
ENSGALG00 000016541	BEND2	114.87 1246	0.6700 5101	0.17868582	3.74988359	0.00017692	0.0085 2385
ENSGALG00 000048948	-	72.294 6449	0.6666 8004	0.24380134	2.73452166	0.0062471	0.0887 8081
ENSGALG00 000043035	SHISA2	13167. 7575	0.6664 6698	0.13178369	5.05727987	4.25E-07	6.91E- 05
ENSGALG00 000040156	CCDC88C	1952.0 935	0.6640 5252	0.1355415	4.89925625	9.62E-07	0.0001 3118
ENSGALG00 000040320	HOXB8	326.05 3234	0.6631 6748	0.21642518	3.06418818	0.00218262	0.0453 9213
ENSGALG00 000002618	KCTD7	565.07 5683	0.6572 0574	0.17741425	3.70435713	0.00021193	0.0096 0251
ENSGALG00 000037506	SIPA1L2	663.88 6969	0.6537 2064	0.18373232	3.55800559	0.00037368	0.0139 4401
ENSGALG00 000030719	CLMN	231.06 9556	0.6416 0906	0.20502943	3.12935105	0.00175193	0.0392 595
ENSGALG00 000011172	MYO10L	369.49 1968	0.6381 8673	0.18627054	3.42612812	0.00061225	0.0199 3127
ENSGALG00 000006503	VDAC1	308.75 4209	0.6350 375	0.23291295	2.72650151	0.00640097	0.0901 7501
ENSGALG00 000033507	PLCG2	128.75 409	0.6318 0985	0.23440756	2.69534762	0.00703152	0.0953 4112
ENSGALG00 000040316	GPC1	2041.1 5525	0.6306 7778	0.12479305	5.05378944	4.33E-07	6.96E- 05
ENSGALG00 000015828	BACH1	412.29 2795	0.6287 1131	0.1641679	3.82968491	0.00012831	0.0066 4715

ENSGALG00 000032028	SERTM1	267.49 9224	0.6279 0934	 0.21454928	 2.92664395	 0.00342641	0.0615 9865
ENSGALG00 000006646	ABCC1	455.47 2921	0.6198 1823	 0.13904785	 4.45758937	 8.29E-06	0.0007 9869
ENSGALG00 000043033	EYA1	1735.6 821	0.6158 6953	 0.19166642	 3.2132365	 0.00131248	0.0329 529
ENSGALG00 000011856	FMC1	336.04 4904	0.6153 8494	 0.18150247	 3.39050454	 0.00069764	0.0217 7906
ENSGALG00 000008909	USP13	98.618 1013	0.6073 1175	 0.21012701	 2.89021275	 0.00384981	0.0664 817
ENSGALG00 000005092	HDAC11	74.419 4314	0.6041 5404	 0.20134055	 3.00065758	 0.00269397	0.0522 667
ENSGALG00 000009028	BTBD3	682.43 0259	0.5902 3334	 0.11522801	 5.12230808	 3.02E-07	5.26E- 05
ENSGALG00 000010313	MAST2	415.43 4906	0.5876 825	 0.21447154	 2.7401422	 0.00614126	0.0875 9364
ENSGALG00 000010885	BMP2K	361.07 088	0.5716 1103	 0.16026548	 3.56665092	 0.00036157	0.0136 8108
ENSGALG00 000013377	LMBRD2	1446.7 0808	0.5643 5284	 0.16780083	 3.36323044	 0.00077036	0.0231 0122
ENSGALG00 000002205	FBRSL1	1447.8 9717	0.5607 1209	 0.12151062	 4.614511	 3.94E-06	0.0004 3909
ENSGALG00 000001730	DGKD	232.54 106	0.5550 878	 0.19618709	 2.82937991	 0.00466383	0.0738 3461
ENSGALG00 000009825	SRGAP1	638.74 7475	0.5546 2103	 0.18594954	 2.98264267	 0.00285771	0.0542 7779
ENSGALG00 000012877	CREB3L2	469.21 9278	0.5497 606	 0.17646374	 3.11543092	 0.00183676	0.0404 6047
ENSGALG00 000005354	ZBTB5	341.19 801	0.5462 0888	 0.19447058	 2.80869674	 0.00497425	0.0768 9602
ENSGALG00 000015132	CDH2	3854.7 4101	0.5458 2154	 0.11725219	 4.6551073	 3.24E-06	0.0003 7084
ENSGALG00 000008238	MED13L	1397.6 5546	0.5379 1441	 0.14844629	 3.62362981	 0.0002905	0.0118 9475
ENSGALG00 000011207	DUSP6	269.30 7863	0.5320 008	 0.16532857	 3.21783956	 0.0012916	0.0327 5227
ENSGALG00 000030802	SHROOM2	521.86 9787	0.5278 656	 0.17108437	 3.08541102	 0.00203271	0.0431 4361

ENSGALG00 000012277	MAFF	186.17 5539	0.5256 2704				0.0329 529
ENSGALG00 000012796	KIAA1549	535.27 3936	0.5251 3841				0.0251 3291
ENSGALG00 000012342	SASH1	780.79 0085	0.5131 8083				0.0387 9915
ENSGALG00 000042340	GTF2H1	375.51 011	0.5128 195				0.0102 0255
ENSGALG00 000046964	CPLANE2	140.10 6167	0.5091 6236				0.0176 2428
ENSGALG00 000042797	-	286.01 1558	0.5062 3686				0.0368 0483
ENSGALG00 000017462	FNIP1	131.14 3002	0.5034 7663				0.0693 8146
ENSGALG00 000052384	-	307.51 9864	0.5000 7055				0.0902 6922
ENSGALG00 000009312	RPL22L1	50988. 7478	0.4999 0345				0.0985 5703
ENSGALG00 000012974	WASH1	124.73 7682	0.4908 7083				0.0496 0761
ENSGALG00 000026600	C1orf198	384.12 2083	0.4903 4158				0.0264 5214
ENSGALG00 000054449	SDHC	272.28 4436	0.4676 1903				0.0872 1751
ENSGALG00 000002180	CHFR	413.89 7585	0.4669 4408				0.0213 2767
ENSGALG00 000040909	WRN	150.73 9824	0.4649 0197				0.0800 2761
ENSGALG00 000043570	SGK3	371.11 934	0.4644 445				0.0365 0714
ENSGALG00 000053766	BOLA3	3201.1 2827	0.4633 8437				0.0043 7181
ENSGALG00 000048612	RPS26	5047.6 2845	0.4607 7647				0.0710 6801
ENSGALG00 000015916	SLC17A5	300.83 0668	0.4579 9282				0.0201 8186
ENSGALG00 000011630	GLI2	4455.9 4996	0.4546 2936				0.0691 0479

ENSGALG00 000014106	PPFIBP1	2278.1 8903	0.4539 649				0.0085 2385
ENSGALG00 000011078	PTPN13	416.24 6284	0.4519 6136				0.0134 289
ENSGALG00 000001491	NCAPD3	389.71 1283	0.4503 1188				0.0417 4583
ENSGALG00 000037375	MAN1B1	236.22 7172	0.4423 3756				0.0638 4523
ENSGALG00 000024391	CACTIN	133.15 2671	0.4370 4932				0.0917 758
ENSGALG00 000031070	MAIP1	2909.3 7141	0.4337 8808				0.0503 9655
ENSGALG00 000011657	EAF2	1191.4 2024	0.4151 1843				0.0408 5094
ENSGALG00 000006856	CD151	567.55 7972	0.4080 4791				0.0262 9502
ENSGALG00 000016813	CHAMP1	985.41 9229	0.4075 7388				0.0299 3824
ENSGALG00 000041914	BRI3	646.56 3799	0.4034 0362				0.0236 1826
ENSGALG00 000007125	PARD3	692.16 4231	0.4018 5094				0.0556 3497
ENSGALG00 000013430	-	369.45 2817	0.3986 4206				0.0893 7156
ENSGALG00 000009021	ST6GALNAC5	312.28 7938	0.3938 3477				0.0902 6922
ENSGALG00 000040651	IGF1R	2424.3 7593	0.3874 6917				0.0512 9524
ENSGALG00 000006426	PAK2	748.75 1659	0.3778 4012				0.0254 8663
ENSGALG00 000038541	VPS37C	534.98 7704	0.3607 5396				0.0634 2553
ENSGALG00 000010743	LRRC42	234.17 4437	0.3519 4263				0.0664 817
ENSGALG00 000028937	TIMM13	4170.0 3567	0.3317 2091				0.0743 1971
ENSGALG00 000006564	DNAJB11	717.17 1328	0.3198 5736				0.0484 4208

ENSGALG0000002606	CLSTN1	1412.3 401	0.3021 1464	0.09811656	3.07914004	0.00207599	0.0438 0492
ENSGALG0000016427	ADAM17	450.53 3213	0.2988 9934	0.10004993	2.98750177	0.00281268	0.0537 0466
ENSGALG0000047617	DDA1	961.79 623	0.2932 1933	0.08723313	3.36132983	0.00077568	0.0232 1263
ENSGALG0000007481	HACD3	1574.3 9876	0.2735 2366	0.09325951	2.93293036	0.00335779	0.0607 2043
ENSGALG0000016463	FAM49A	1850.6 2033	0.2364 8944	0.08684236	2.72320377	0.00646522	0.0905 5062

Table 3.2.9: List of 190 downregulated genes in thoracic somites vs. cervical somites from the overlap of D1 and D2, excluding D3, from the RNA-Seq dataset

Gene ID	Gene Name	Base Mean	Log2 Fold Change	lfcSE	stat	pvalue	padj
ENSGALG0000011270	KERA	35.908 9176	-6.8193374	1.10571484	- 6.1673564	6.94E-10	3.58E-07
ENSGALG0000003478	GFI1B	3.5263 0748	-4.956925	1.43574133	- 3.4525196	0.00055538	0.0187 1194
ENSGALG0000015148	ANXA1	27.492 4078	-4.5888888	0.69454833	- 6.6070115	3.92E-11	2.83E-08
ENSGALG0000002263	KLHL6	5.5338 5801	-4.5455909	1.30730884	- 3.4770597	0.00050695	0.0176 139
ENSGALG0000034806	SLC7A7	6.8028 7589	-4.3950697	0.95254291	- 4.6140386	3.95E-06	0.0004 3909
ENSGALG0000036142	LOXL3	35.805 1675	-4.249731	0.83282571	- 5.1027856	3.35E-07	5.56E-05
ENSGALG0000045520	-	5.6110 8033	-3.2997759	1.10094589	-2.997219	0.00272455	0.0525 0752
ENSGALG0000047483	C1QTNF6	16.395 2902	-3.1712466	0.67046843	-4.729897	2.25E-06	0.0002 6834
ENSGALG0000005441	NFIB	6.8391 9651	-3.1676486	0.92034794	- 3.4417946	0.00057787	0.0191 9109
ENSGALG0000039040	PAX1	36.247 9835	-3.1280462	0.81976352	-3.815791	0.00013575	0.0069 0878
ENSGALG0000039452	GJA5	512.14 7159	-2.6046571	0.74885329	- 3.4781942	0.0005048	0.0175 8178

ENSGALG00 000004290	ANGPTL1	151.04 6272	-2.2547177	0.44958509	- 5.0151078	5.30E-07	8.24E- 05
ENSGALG00 000011473	RPS6KA2	13.710 2118	-2.2033764	0.72029889	- 3.0589751	0.00222096	0.0460 5696
ENSGALG00 000011902	TAGLN	10.258 2429	-2.1278901	0.63703572	- 3.3402995	0.00083688	0.0244 2793
ENSGALG00 000013409	-	14.399 7467	-2.1222208	0.59562315	- 3.5630261	0.0003666	0.0137 6337
ENSGALG00 000005352	GATA5	75.010 2984	-2.0022745	0.73690436	-2.717143	0.00658482	0.0914 2239
ENSGALG00 000009594	RARRES1	656.68 3732	-1.9884689	0.44926555	- 4.4260435	9.60E-06	0.0008 8697
ENSGALG00 000001136	-	102.53 6116	-1.9527095	0.6186888	- 3.1562063	0.00159836	0.0372 6236
ENSGALG00 000054209	RGS3	78.822 2177	-1.9407731	0.53624994	- 3.6191577	0.00029556	0.0120 3401
ENSGALG00 000011306	DES	37.910 5275	-1.9397849	0.41529452	- 4.6708657	3.00E-06	0.0003 4682
ENSGALG00 000011855	RHOJ	36.300 8598	-1.9149082	0.5727238	- 3.3435108	0.00082726	0.0242 5383
ENSGALG00 000009512	TFPI2	101.21 2299	-1.7539327	0.37547797	-4.6712	2.99E-06	0.0003 4682
ENSGALG00 000008007	ITGA11	16.681 2302	-1.7207289	0.62790199	- 2.7404419	0.00613566	0.0875 9364
ENSGALG00 000054680	-	225.16 7881	-1.7114095	0.32732335	-5.228498	1.71E-07	3.29E- 05
ENSGALG00 000011450	LTC4SL	24.030 7906	-1.631102	0.57755456	- 2.8241522	0.00474059	0.0743 9788
ENSGALG00 000003939	KLF2	10.929 9203	-1.5668617	0.57508732	- 2.7245631	0.00643866	0.0903 9332
ENSGALG00 000002576	ASL2	15.017 2596	-1.4305737	0.45394395	- 3.1514324	0.00162472	0.0375 7387
ENSGALG00 000000433	TUBA4A	76.623 3894	-1.4069689	0.3357987	- 4.1899175	2.79E-05	0.0021 5694
ENSGALG00 000049587	-	26.578 1768	-1.3791651	0.39959643	-3.451395	0.0005577	0.0187 4639
ENSGALG00 000000379	ITGB3	61.903 3073	-1.343783	0.48539775	- 2.7684163	0.00563295	0.0835 7277

ENSGALG00 000001956	OLFML3	22449. 6326	-1.2243778	0.22296739	- 5.4912863	3.99E-08	9.61E- 06
ENSGALG00 000023437	NQO1	44.046 5255	-1.2063128	0.42355854	- 2.8480427	0.0043989	0.0715 2051
ENSGALG00 000013297	BBOX1	673.92 8407	-1.2043363	0.19689121	- 6.1167601	9.55E-10	4.60E- 07
ENSGALG00 000042020	CG-16	344.31 7026	-1.2004584	0.30026994	- 3.9979306	6.39E-05	0.0039 1578
ENSGALG00 000008546	FGD5	651.94 5516	-1.1964786	0.34156643	- 3.5029163	0.00046019	0.0163 4311
ENSGALG00 000010854	-	74.813 0408	-1.1790773	0.31603167	- 3.7308833	0.00019081	0.0089 5442
ENSGALG00 000011324	TUBA4B	256.51 7247	-1.1690656	0.30602338	- 3.8201839	0.00013335	0.0068 1086
ENSGALG00 000032588	ARPC1B	93.727 147	-1.1666566	0.30344012	- 3.8447672	0.00012067	0.0063 654
ENSGALG00 000030647	MFAP2	865.10 7739	-1.150826	0.20109614	- 5.7227651	1.05E-08	3.44E- 06
ENSGALG00 000008206	TESC	527.54 8538	-1.1486401	0.23886784	- 4.8086848	1.52E-06	0.0001 8931
ENSGALG00 000010794	MEOX2	231.08 7367	-1.1433727	0.26809654	-4.26478	2.00E-05	0.0016 4332
ENSGALG00 000030209	CPZ	2725.8 2044	-1.1084745	0.28281926	- 3.9193743	8.88E-05	0.0049 9305
ENSGALG00 000043499	EGFL7	394.18 2236	-1.1001529	0.29659922	- 3.7092239	0.0002079	0.0094 4944
ENSGALG00 000005503	WFDC1	218.12 3721	-1.0780641	0.21645392	- 4.9805709	6.34E-07	9.45E- 05
ENSGALG00 000018702	-	55.475 1361	-1.0572056	0.33889902	-3.11953	0.0018114	0.0399 7245
ENSGALG00 000004386	SOX9	109.68 7721	-1.0359548	0.34714061	- 2.9842512	0.00284273	0.0541 3552
ENSGALG00 000004028	MATN4	93.109 0101	-1.0349256	0.35881675	- 2.8842734	0.00392318	0.0671 0727
ENSGALG00 000003589	VTN	88.514 9117	-1.0236849	0.27649031	-3.702426	0.00021355	0.0096 4568
ENSGALG00 000014686	FBN2	200.89 2641	-1.0180986	0.29371868	- 3.4662372	0.0005278	0.0180 7769

ENSGALG00 000016328	GSTA2	849.36 4859	-0.9943083	0.21530121	- 4.6182197	3.87E-06	0.0004 3706
ENSGALG00 000006546	CD81	1224.1 7404	-0.9284455	0.1995016	- 4.6538248	3.26E-06	0.0003 7084
ENSGALG00 000009641	COL1A2	11355. 9119	-0.892154	0.22942522	- 3.8886482	0.0001008	0.0055 4001
ENSGALG00 000040298	TPK1	285.57 6519	-0.8914006	0.169654	- 5.2542267	1.49E-07	3.01E- 05
ENSGALG00 000030376	DHRS3	202.75 1476	-0.8856469	0.17638295	- 5.0211594	5.14E-07	8.07E- 05
ENSGALG00 000006405	CEP19	78.670 1615	-0.8822933	0.24970853	- 3.5332926	0.00041042	0.0149 0498
ENSGALG00 000002771	CRIP1	1465.9 7961	-0.8822637	0.31777961	- 2.7763384	0.0054975	0.0820 0289
ENSGALG00 000000785	HYAL1	110.67 4186	-0.8621967	0.19755214	- 4.3644006	1.27E-05	0.0011 3035
ENSGALG00 000017048	EXOSC8	2013.2 2146	-0.8564843	0.22058073	- 3.8828607	0.00010324	0.0056 0663
ENSGALG00 000016322	-	113.67 9689	-0.8499256	0.19384489	- 4.3845653	1.16E-05	0.0010 4988
ENSGALG00 000010034	-	995.95 7915	-0.8206252	0.18866333	- 4.3496807	1.36E-05	0.0011 943
ENSGALG00 000006250	TRAPPC2L	2859.7 2861	-0.8184042	0.23808746	-3.43741	0.00058731	0.0194 2544
ENSGALG00 000011298	RARB	343.46 4139	-0.8024453	0.18423492	- 4.3555546	1.33E-05	0.0011 6981
ENSGALG00 000033462	-	234.65 1305	-0.7986691	0.29453468	- 2.7116299	0.00669533	0.0922 5386
ENSGALG00 000000632	CHTF8	2477.6 3297	-0.7975741	0.23133297	- 3.4477322	0.00056531	0.0189 5219
ENSGALG00 000003636	FBLIM1	157.45 3375	-0.7916043	0.22386991	- 3.5360016	0.00040623	0.0147 9013
ENSGALG00 000003123	CNTNAP1	119.84 7837	-0.7836821	0.24669264	-3.176755	0.00148933	0.0358 1821
ENSGALG00 000041640	TWIST1	507.13 0421	-0.7779925	0.19404388	- 4.0093635	6.09E-05	0.0037 6347
ENSGALG00 000008018	SNAI1	3289.8 2692	-0.7779699	0.21707079	- 3.5839458	0.00033844	0.0132 9306

ENSGALG00 000024041	FAHD1	207.28 1133	-0.7660291	0.20889599	-3.667036	0.00024538	0.0106 1327
ENSGALG00 000006409	PODXL	1137.7 3147	-0.7459302	0.15129267	- 4.9303787	8.21E-07	0.0001 1632
ENSGALG00 000027860	CISD2	4264.9 8877	-0.7395544	0.23540101	-3.141679	0.00167982	0.0383 5723
ENSGALG00 000005744	EMC8	5833.3 1655	-0.737556	0.21203186	- 3.4785149	0.0005042	0.0175 8178
ENSGALG00 000005422	PDLIM1	113.95 9286	-0.7342531	0.25338094	- 2.8978229	0.00375763	0.0656 647
ENSGALG00 000047563	-	28.375 5019	-0.7340585	0.26619496	- 2.7575971	0.00582279	0.0850 729
ENSGALG00 000016932	MZT1	1660.9 3002	-0.732257	0.23333433	- 3.1382309	0.00169971	0.0386 128
ENSGALG00 000014708	GSTK1	281.23 2826	-0.717785	0.16212089	- 4.4274678	9.53E-06	0.0008 8697
ENSGALG00 000010052	PPP2R3C	419.55 8399	-0.7081422	0.20638292	- 3.4312052	0.00060091	0.0197 7996
ENSGALG00 000010157	TRAPPC6A	872.50 5265	-0.7014983	0.20050987	- 3.4985724	0.00046776	0.0165 1259
ENSGALG00 000021225	CDCA9	649.25 7948	-0.6857296	0.22209802	- 3.0875089	0.00201842	0.0431 3981
ENSGALG00 000004593	SULF2	255.66 3223	-0.6849917	0.21859561	- 3.1336024	0.00172675	0.0389 3665
ENSGALG00 000032400	UTP23	625.09 0152	-0.6818986	0.20335014	- 3.3533227	0.00079848	0.0236 1826
ENSGALG00 000026385	LYPLAL1	291.01 2234	-0.6789254	0.2036753	- 3.3333712	0.000858	0.0249 0279
ENSGALG00 000001143	ETS1	414.81 5012	-0.6780678	0.2275525	- 2.9798302	0.00288408	0.0545 6351
ENSGALG00 000004509	-	186005 .087	-0.6745105	0.18820039	- 3.5840014	0.00033837	0.0132 9306
ENSGALG00 000013018	TPRKB	908.81 6737	-0.6692941	0.19540283	- 3.4252015	0.00061434	0.0199 4945
ENSGALG00 000000629	-	1361.2 0968	-0.6691174	0.20395494	- 3.2807119	0.00103546	0.0282 2877
ENSGALG00 000040360	DUSP1	130.37 9009	-0.6661563	0.18604752	- 3.5805709	0.00034284	0.0133 6094

ENSGALG00 000035369	-	148.77 7189	-0.656804	0.20453577	- 3.2111937	0.00132185	0.0329 529
ENSGALG00 000039387	PSMA2	21469. 4629	-0.6554461	0.17251186	- 3.7994264	0.00014503	0.0071 5455
ENSGALG00 000011293	MRPL42	1795.6 232	-0.6553297	0.19254878	- 3.4034476	0.00066541	0.0210 585
ENSGALG00 000011386	BCS1L	386.87 8498	-0.6439692	0.21826513	-2.950399	0.00317364	0.0584 3536
ENSGALG00 000004057	HMGCL	376.76 5305	-0.6393839	0.20223954	- 3.1615177	0.00156949	0.0369 6965
ENSGALG00 000019036	SPCS2	624.28 9806	-0.6372677	0.21069837	- 3.0245497	0.00249004	0.0495 0615
ENSGALG00 000014261	UCHL1	2047.4 4802	-0.6364484	0.18334311	- 3.4713513	0.00051785	0.0178 213
ENSGALG00 000007190	ATP6V1G1	22312. 2843	-0.6325585	0.20030167	-3.158029	0.0015884	0.0371 3677
ENSGALG00 000005229	DPH5	770.71 8822	-0.6288819	0.180963	- 3.4751961	0.00051048	0.0176 2428
ENSGALG00 000001472	JAM3	2521.6 4116	-0.6255315	0.15263918	-4.098106	4.17E-05	0.0029 2269
ENSGALG00 000036214	PARK7	4012.9 3403	-0.6217056	0.1923189	- 3.2326806	0.00122635	0.0315 9645
ENSGALG00 000002587	CRCP	1350.6 0294	-0.6195858	0.1497764	- 4.1367386	3.52E-05	0.0026 1119
ENSGALG00 000008538	GIN51	441.14 8509	-0.6164992	0.15861989	- 3.8866451	0.00010164	0.0055 6475
ENSGALG00 000038812	MRPL27	4700.5 8243	-0.6110081	0.19403145	- 3.1490161	0.00163821	0.0376 4501
ENSGALG00 000008291	AP3S2	627.67 6736	-0.6079542	0.17864998	- 3.4030468	0.00066639	0.0210 585
ENSGALG00 000033245	SMPD2	129.96 6478	-0.6006957	0.19446761	-3.088924	0.00200883	0.0431 3981
ENSGALG00 000026383	TMSB4X	45596. 8075	-0.6000953	0.19290894	- 3.1107701	0.001866	0.0408 5094
ENSGALG00 000040167	TPD52	340.99 3324	-0.5999597	0.18748691	- 3.2000084	0.00137424	0.0336 6645
ENSGALG00 000037067	EIF3E	6757.6 2165	-0.5998895	0.20224673	- 2.9661274	0.00301576	0.0563 1748

ENSGALG00 000016920	LMO7	142.81 7646	-0.5982661	0.20733515	- 2.8855026	0.00390789	0.0670 0436
ENSGALG00 000010285	MRPL14	2798.6 0423	-0.5963877	0.18855552	- 3.1629287	0.00156191	0.0369 4892
ENSGALG00 000043733	TMEM50A	1076.2 7431	-0.5931071	0.19003361	- 3.1210645	0.00180199	0.0398 5478
ENSGALG00 000002379	MRPS17	3768.8 0689	-0.5910248	0.19859073	- 2.9760947	0.00291945	0.0550 8837
ENSGALG00 000014582	CADM1	1837.0 3502	-0.5900933	0.14318916	- 4.1210753	3.77E-05	0.0027 3517
ENSGALG00 000017291	FAR2	414.42 6577	-0.5853095	0.19518716	-2.998709	0.00271126	0.0523 212
ENSGALG00 000030200	NFU1	1596.5 2849	-0.5845373	0.19373332	- 3.0172264	0.00255099	0.0503 5545
ENSGALG00 000004705	BUD31	8784.3 8344	-0.5819168	0.18282256	- 3.1829593	0.00145778	0.0353 5062
ENSGALG00 000013569	SEC61B	22725. 2028	-0.5792505	0.19885235	-2.912968	0.00358011	0.0637 0206
ENSGALG00 000014887	ASF1A	4645.5 1288	-0.5780338	0.175539	- 3.2929079	0.00099157	0.0276 1887
ENSGALG00 000010569	TXNDC12	1219.1 5526	-0.5780144	0.14674464	- 3.9389135	8.19E-05	0.0047 1347
ENSGALG00 000011953	PHF5A	7450.9 8189	-0.5780074	0.20165835	- 2.8662707	0.00415339	0.0692 5593
ENSGALG00 000051398	-	561.00 5075	-0.5764985	0.19864338	- 2.9021781	0.00370578	0.0650 8863
ENSGALG00 000041972	POP4	685.07 0989	-0.5747584	0.20470797	- 2.8076995	0.00498968	0.0770 5212
ENSGALG00 000006022	PSMA1	16120. 059	-0.5741134	0.17853928	- 3.2156141	0.00130166	0.0328 9188
ENSGALG00 000017366	DPM1	4175.3 761	-0.5709589	0.16141059	- 3.5373076	0.00040423	0.0147 7443
ENSGALG00 000026271	TM2D3	366.12 1791	-0.5705991	0.17356156	-3.287589	0.00101049	0.0278 7339
ENSGALG00 000003750	PLCG1	1853.1 3749	-0.5678884	0.12668731	- 4.4825987	7.37E-06	0.0007 3002
ENSGALG00 000012379	CSH14orf166	13578. 2331	-0.565717	0.18533935	-3.052331	0.00227072	0.0467 5345

ENSGALG00 000015808	SLC35A1	762.40 3003	-0.5628116	0.15876707	- 3.5448886	0.00039278	0.0144 8273
ENSGALG00 000028354	METRNL	657.92 0724	-0.5561419	0.20807658	- 2.6727749	0.00752267	0.0988 7507
ENSGALG00 000013759	MRPL47	1509.4 2085	-0.5560596	0.1537435	-3.616801	0.00029827	0.0120 8234
ENSGALG00 000040878	TMEM208	1058.3 4806	-0.5558127	0.18647996	- 2.9805493	0.00287732	0.0545 0691
ENSGALG00 000001330	ATP5MC1	3989.0 0844	-0.5519765	0.19055138	- 2.8967334	0.0037707	0.0656 647
ENSGALG00 000036288	OAZ2	3079.1 0318	-0.5480607	0.17961449	- 3.0513168	0.0022784	0.0468 4495
ENSGALG00 000034396	TMEM50B	178.77 4428	-0.5476277	0.19024544	- 2.8785327	0.0039953	0.0677 794
ENSGALG00 000023576	PIM3	330.10 2555	-0.5408405	0.17699824	- 3.0556263	0.00224591	0.0464 8871
ENSGALG00 000009993	MED8	733.56 1423	-0.5394563	0.18154547	- 2.9714666	0.00296381	0.0556 3497
ENSGALG00 000040987	NDUFA8	10338. 3336	-0.5375244	0.18030288	- 2.9812305	0.00287093	0.0544 5718
ENSGALG00 000021665	-	105.70 498	-0.5325517	0.19005214	-2.802135	0.00507656	0.0781 4339
ENSGALG00 000027070	TIMP2	317.17 7255	-0.531167	0.17280716	- 3.0737557	0.00211383	0.0443 443
ENSGALG00 000012087	CEP70	305.10 1061	-0.528806	0.1354399	- 3.9043588	9.45E-05	0.0052 5211
ENSGALG00 000007473	MRPL28	5350.1 2997	-0.528734	0.18669783	- 2.8320307	0.00462534	0.0736 9663
ENSGALG00 000053063	MRPL40	1835.3 8376	-0.5274261	0.16619121	- 3.1736101	0.00150556	0.0361 484
ENSGALG00 000036218	RAB4A	820.79 623	-0.5257162	0.1859617	- 2.8270136	0.00469843	0.0742 1984
ENSGALG00 000008818	IDH1	644.03 3021	-0.5222537	0.11603514	- 4.5008238	6.77E-06	0.0006 8445
ENSGALG00 000028290	C1H21ORF91	895.83 8586	-0.5212882	0.17120101	- 3.0448897	0.00232766	0.0475 8692
ENSGALG00 000011689	PDIA5	2266.7 4936	-0.5073914	0.14431255	- 3.5159204	0.00043823	0.0157 1765

ENSGALG00 000004799	-	194.02 6207	-0.5018988	0.17756629	- 2.8265433	0.00470534	0.0742 4777
ENSGALG00 000008506	MRPS25	2315.6 2591	-0.4992192	0.18678583	- 2.6726827	0.00752474	0.0988 7507
ENSGALG00 000001729	CD276	2466.6 655	-0.4969714	0.14485031	- 3.4309307	0.00060151	0.0197 7996
ENSGALG00 000013235	PDK1	343.34 0743	-0.4924662	0.15577773	- 3.1613391	0.00157046	0.0369 6965
ENSGALG00 000040511	MTERF3	2483.2 6143	-0.491611	0.15697573	- 3.1317642	0.00173759	0.0390 289
ENSGALG00 000006753	ATP5C1	31916. 1656	-0.4907662	0.17544694	- 2.7972346	0.00515421	0.0785 8539
ENSGALG00 000007822	MRPL58	1245.8 1991	-0.4906209	0.17839814	- 2.7501459	0.00595687	0.0860 2389
ENSGALG00 000019719	KRT24	144.83 2943	-0.4899146	0.17290875	- 2.8333708	0.00460599	0.0736 9663
ENSGALG00 000041277	UBB	227021 .557	-0.487315	0.14520664	- 3.3560109	0.00079076	0.0235 1763
ENSGALG00 000005132	CIAO2B	1273.1 8418	-0.487099	0.16560088	- 2.9414032	0.00326729	0.0596 6678
ENSGALG00 000046258	FAM92A	243.34 356	-0.486065	0.1498048	- 3.2446559	0.00117593	0.0307 357
ENSGALG00 000014023	H2AFZ	69603. 5348	-0.4853407	0.1714534	- -2.830744	0.00464399	0.0736 9663
ENSGALG00 000006132	MOSPD1	1143.4 3041	-0.4815359	0.16460521	- 2.9253988	0.00344015	0.0617 6885
ENSGALG00 000003549	MTG1	889.96 4643	-0.4743285	0.17315203	- 2.7393759	0.00615559	0.0876 5809
ENSGALG00 000049287	ATP6V0E1	811.34 5676	-0.4673842	0.17376033	- 2.6898214	0.00714903	0.0963 0198
ENSGALG00 000017129	SAP18	24415. 4481	-0.4664479	0.17343968	- 2.6893957	0.00715815	0.0963 3511
ENSGALG00 000038604	PHB	8044.2 8785	-0.4662155	0.16049016	- 2.9049474	0.00367315	0.0649 0426
ENSGALG00 000007145	ITGB1	3364.1 5655	-0.4656566	0.11460065	- 4.0632975	4.84E-05	0.0031 7885
ENSGALG00 000005205	RNFT1	151.55 6384	-0.4648831	0.16100118	- 2.8874518	0.00388376	0.0667 4896

ENSGALG00 000003991	MRT04	2753.5 5294	-0.4626841	0.16824199	- 2.7501106	0.00595751	0.0860 2389
ENSGALG00 000026031	ZNHIT3	2056.2 9385	-0.4586453	0.17122204	- 2.6786579	0.00739179	0.0981 0918
ENSGALG00 000016991	ESD	1314.6 8855	-0.4479712	0.1330138	- 3.3678551	0.00075755	0.0229 5531
ENSGALG00 000042023	-	608.02 5236	-0.4472713	0.14949877	- 2.9918056	0.00277333	0.0530 2339
ENSGALG00 000001792	MRPS34	834.97 1503	-0.4462865	0.11106325	- 4.0183092	5.86E-05	0.0036 998
ENSGALG00 000008858	MDH1	9924.3 4933	-0.4405355	0.11851483	- 3.7171339	0.0002015	0.0093 0326
ENSGALG00 000049929	LIAS	417.53 8981	-0.4399326	0.1303331	- 3.3754482	0.00073696	0.0226 6374
ENSGALG00 000015436	SDCBP	7728.9 8845	-0.4307261	0.14177136	- 3.0381743	0.00238016	0.0483 4712
ENSGALG00 000039171	RAD52	343.43 1591	-0.4266849	0.12304821	- 3.4676239	0.00052508	0.0180 2739
ENSGALG00 000017282	NDUFA9	5701.9 7953	-0.4144898	0.14584867	- 2.8419168	0.00448432	0.0723 3968
ENSGALG00 000001547	CAPZA1	4175.7 3393	-0.3999511	0.12413259	- 3.2219668	0.00127314	0.0325 2918
ENSGALG00 000036730	MRPS35	1349.9 5779	-0.3983272	0.14605953	- 2.7271565	0.00638827	0.0900 8399
ENSGALG00 000010464	GLRX3	4532.0 756	-0.3917335	0.13639737	- 2.8720019	0.00407881	0.0686 3218
ENSGALG00 000004440	VPS33A	429.16 6045	-0.390261	0.14150384	- 2.7579531	0.00581645	0.0850 729
ENSGALG00 000012387	NUP43	2682.7 1393	-0.3810379	0.13209854	- 2.8844972	0.00392039	0.0671 0727
ENSGALG00 000005894	USP8	279.23 0822	-0.3798517	0.13970105	- 2.7190325	0.00654732	0.0912 5837
ENSGALG00 000019976	ADAT2	483.55 4727	-0.3750842	0.14008873	- 2.6774758	0.00741792	0.0983 6573
ENSGALG00 000001962	PTBP1	9626.0 9385	-0.3439241	0.11145566	-3.085748	0.00203041	0.0431 4361
ENSGALG00 000002094	TRMT2A	131.15 9086	-0.3417197	0.12251539	- 2.7891986	0.00528387	0.0798 8807

ENSGALG00 000004531	PLOD1	794.89 6195	-0.3318686	0.11942625	- 2.7788581	0.00545504	0.0815 3782
ENSGALG00 000011318	DNPEP	2399.8 4643	-0.3312601	0.11227342	- 2.9504763	0.00317284	0.0584 3536
ENSGALG00 000030843	COPSS	1088.4 103	-0.318914	0.10360954	- 3.0780372	0.00208369	0.0439 0326
ENSGALG00 000016736	CYFIP1	2087.5 8124	-0.3094364	0.11055158	- 2.7990228	0.00512575	0.0784 8264
ENSGALG00 000007757	CSTF1	505.36 3242	-0.3078671	0.10612329	- 2.9010319	0.00371936	0.0652 4228
ENSGALG00 000003265	PTP4A2	1462.3 0229	-0.2561956	0.09145386	- 2.8013648	0.0050887	0.0782 4682
ENSGALG00 000008504	HNRNPLL	1222.8 0383	-0.254635	0.0878673	- 2.8979494	0.00375611	0.0656 647
ENSGALG00 000023682	GTF2H3	1341.5 2504	-0.2526106	0.09386358	- 2.6912524	0.00711843	0.0962 4865

3.1. Functions used in R Programme

Table 3.1.10: R packages used to generate the RNA-Seq dataset and for the subsequent analyses throughout.

R Package	Use
ggplot2	a system for declaratively creating graphics
BiocParallel	provides modified versions and novel implementation of functions for parallel evaluation
DESeq2	estimates variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution
pheatmap	a function to draw clustered heatmaps
gplots	various R programming tools for plotting data
RColorBrewer	contains a ready-to-use color palettes for creating beautiful graphics
sva	contains functions for removing batch effects and other unwanted variation in high-throughput experiment
pamr	prediction analysis for microarrays
limma	differential expression analyses for RNA-sequencing and microarray studies

The R function below was used to normalise the RNA-Seq dataset.

```
countsTable<-read.delim("expression_matrix.txt", header=T, row.names=1)
countData<-matrix(countsTable)
dim(countData)
sampleInfo<-read.table("annotation.txt", header=T)
attach(sampleInfo)

sampleInfo

meta<-data.frame(row.names=colnames(countsTable), cell_type=factor(Type), sampe_names=factor(Name))
dds<-DESeqDataSetFromMatrix(countData=countsTable,meta,formula(~cell_type))
idx <- which(rowSums(counts(dds)) > 0)
dds <- dds[idx,]
dds <- estimateSizeFactors(dds)

dds2 <- DESeq(dds, parallel=T)
res <- results(dds2)

summary(res)

write.table(res, file="DE.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)

temp2<-counts(dds, normalized=TRUE)

write.table(temp2, file="normalized_counts.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)
```

The R function below was used to generate the various PCA plots for analysis.

```
vsd <- vst(dds, blind=FALSE)
write.table(assay(vsd), file="vsd_counts.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)
sampleDist<-dist(t(assay(vsd)))
sampleDistMatrix <- as.matrix(sampleDist)
rownames(sampleDistMatrix) <- paste(vsd$position)
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues"))) (255)
pheatmap(sampleDistMatrix, clustering_distance_rows=sampleDist, clustering_distance_cols=sampleDist, col=colors, treeheight_row=0, fontsize=5)

pcaResult<-prcomp(t(assay(vsd)))
summary(pcaResult)

write.table(pcaResult$x, file="pca_results.txt", sep="\t", quote=F, col.names = TRUE, row.names = TRUE)

head(pcaResult$x)
#coordinates of the samples on the different components

temp2<-pcaResult$rotation
head(temp2)

write.table(pcaResult$rotation, file="pca_results_genes.txt", sep="\t", quote=F, col.names = TRUE, row.names = TRUE)

temp<-read.table("pca_results_sum_short.txt", header=T)
attach(temp)

cor.test(HOX, PC1, method="spearman")
cor.test(SUM, PC2, method="spearman")
cor.test(read_count, PC3, method="spearman")

p<-ggplot(temp, aes(PC1, PC2, color=REP))+theme_bw()+geom_point()+ylab("PC2 (15.92%)")+xlab("PC1 (21.33%)")

p<-ggplot(temp, aes(PC1, PC2, color=log10(SUM+1)))+theme_bw()+geom_point()+ylab("PC2 (15.92%)")+xlab("PC1 (21.33%)")+ scale_colour_gradient(low = "blue4", high = "red")
```


The R function below was used for batch removal from the RNA-Seq dataset.

```
batch <- c(rep(1, 12), rep(2, 12))
batch
countsTable<-read.delim("expression_matrix.txt", header=T, row.names=1)
countData<-as.matrix(countsTable)
adjusted <- ComBat_seq(countData, batch=batch, group=NULL)
head(adjusted)

write.table(adjusted, file="adjusted_matrix_counts.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)

countsTable<-read.delim("adjusted_matrix_counts.txt", header=T, row.names=1)
countData<-matrix(countsTable)
dim(countData)
sampleInfo2<-read.table("annotation.txt", header=T)
attach(sampleInfo2)
head(sampleInfo2)
meta<-data.frame(row.names=colnames(countsTable), cell_type=factor(Type), sampe_names=factor(Name))
dds<-DESeqDataSetFromMatrix(countData=countData,meta,formula(~cell_type))
idx <- which(rowSums(counts(dds)) > 0)
dds <- dds[idx,]
dds <- estimateSizeFactors(dds)

dds2 <- DESeq(dds, parallel=T)
res <- results(dds2)

summary(res)

write.table(res, file="DE_short_adjusted.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)

temp2<-counts(dds, normalized=TRUE)

write.table(temp2, file="normalized_counts_short_adjusted.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)
```

The R function below was used to generate PCA plots following batch removal.

```
vsd <- vst(dds, blind=FALSE)
sampleDist<-dist(t(assay(vsd)))
sampleDistMatrix <- as.matrix(sampleDist)
rownames(sampleDistMatrix) <- paste(vsd$position)
colnames(sampleDistMatrix) <- NULL
colors <- colorRampPalette( rev(brewer.pal(9, "Blues"))) )(255)
pheatmap(sampleDistMatrix, clustering_distance_rows=sampleDist, clustering_distance_cols=sampleDist, col=colors, treeheight_row=0, fontsize=5)

pcaResult<-prcomp(t(assay(vsd)))
summary(pcaResult)
write.table(pcaResult$x, file="pca_results_short_adjusted.txt", sep="\t", quote=F, col.names = TRUE, row.names = TRUE)

temp<-read.table("pca_results_short_adjusted.txt", header=T)
attach(temp)

head(temp)

p<-ggplot(temp, aes(PC1, PC2, color=somite))+theme_bw()+geom_point()+ylab("PC2 (10.20%)")+xlab("PC1 (20.42%)")
p
```

The R function below was used to generate MA plots and heatmaps.

```
plotMA(res, ylim=c(-2,2))
plotCounts(dds2, gene=which.min(res$padj), intgroup = "cell_type")
plotCounts(dds2, gene="ENSGALG00000039051", intgroup = "cell_type")
plotPCA(vsd, intgroup=c("cell_type"))
plotPCA(vsd, intgroup=c("sampe_names"))
library("genefilter")
topVarGenes <- head(order(-rowVars(assay(vsd))),100)
mat <- assay(vsd)[ topVarGenes, ]
mat <- mat - rowMeans(mat)
df <- as.data.frame(colData(vsd)[,c("cell_type", "sampe_names")])
pheatmap(mat, annotation_col=df)
pheatmap(mat,
          annotation_col = df,
          cutree_rows = 2,
          cutree_cols = 2)
out = pheatmap(mat,
               show_rownames = F, cluster_cols=T, cluster_rows=T, scale="row",
               cex=1, clustering_distance_rows = "euclidean", cex=1,
               clustering_distance_cols = "euclidean", clustering_method = "complete", border_color = FALSE)
res <- mat[c(out$tree_row[["order"]]], out$tree_col[["order"]])
res
write.table(res, file="pheatmapgenes200.txt", quote=F, row.names = TRUE, col.names = TRUE)
write.table(temp2, file="pheatmapgenes200.txt", sep="\t", quote=F, row.names = TRUE, col.names = TRUE)
```

The R function below was used to carry out hierarchical clustering of the RNA-Seq dataset.

```
trans_cts<-read.delim("normalized_counts_SW_newfinal_cerv2vpost_label.txt", header=T)
sample_info<-read.table("SW_annotation_cerv2vpost.txt", header=T)
test_result<-read.delim("DE_newfinal_cerv2vpost_SW copy.txt", header=T)

head(test_result$padj)

candidate_genes <- test_result %>%
  filter(test_result$padj < 0.01) %>%
  pull(gene) %>%
  unique()
head(candidate_genes)

trans_cts_mean <- trans_cts %>%
  # convert to long format
  pivot_longer(cols = D41:D71, names_to = "Sample", values_to = "cts") %>%
  # join with sample info table
  full_join(sample_info, by = ("Sample")) %>%
  # filter to retain only genes of interest
  filter(gene %in% candidate_genes) %>%
  # for each gene
  group_by(gene) %>%
  # scale the cts column
  mutate(cts_scaled = (cts - mean(cts))/sd(cts)) %>%
  # for each gene, strain and minute
  group_by(gene, somite) %>%
  # calculate the mean (scaled) cts
  summarise(mean_cts_scaled = mean(cts_scaled),
            nrep = n()) %>%
  ungroup()

hclust_matrix <- trans_cts %>%
  select(-gene) %>%
  as.matrix()

rownames(hclust_matrix) <- trans_cts$gene
```

```

hclust_matrix <- trans_cts %>%
  select(-gene) %>%
  as.matrix()

rownames(hclust_matrix) <- trans_cts$gene

is.list(candidate_genes)
hclust_matrix <- hclust_matrix[candidate_genes, ]

hclust_matrix <- hclust_matrix %>%
  t() %>%
  scale() %>%
  t()

gene_dist <- dist(hclust_matrix)

gene_hclust <- hclust(gene_dist, method = "complete")

plot(gene_hclust, labels = FALSE)
abline(h = 7, col = "brown", lwd = 2)

cutree(gene_hclust, k = 22)

gene_cluster$name

gene_cluster <- cutree(gene_hclust, k = 22) %>%
  # turn the named vector into a tibble
  enframe() %>%
  # rename some of the columns
  rename(gene = name, cluster = value)

```

```

head(gene_cluster)
write.table(gene_cluster, file = "clusterugh.txt", sep = "\t", row.names = TRUE, col.names = TRUE)

trans_cts_cluster <- trans_cts_mean %>%
  inner_join(gene_cluster, by = "gene")

head(trans_cts_cluster)

trans_cts_cluster %>%
  ggplot(aes(somite, mean_cts_scaled)) +
  geom_line(aes(group = gene)) +
  facet_grid(cols = vars(cluster))

trans_cts_cluster %>%
  ggplot(aes(somite, mean_cts_scaled)) +
  geom_line(aes(group = gene), alpha = 0.3) +
  geom_line(stat = "summary", fun.y = "median", colour = "brown", size = 1.5,
    aes(group = 1)) +
  facet_grid(cols = vars(cluster))

library(ComplexHeatmap)
Heatmap(hclust_matrix, show_row_names = FALSE)

```